Urednik
ZLATKO ŠPORER

Recenzenti
Dr. *SREĆKO POLIĆ*
Dr. *NIKOLA SARAPA*
Dr. *IVAN ŠOŠIĆ*
Dr. *DIMITRIJE UGRIN ŠPARAC*

Lektorica
GIOIA VUČINIĆ

Grafički urednik
ŽELJKO IVANČIĆ

Korektorica
BRANKA MESING-NAKARADA

Slog i prijelom: *T & S, Zagreb*

Objavljivanje ovog sveučilišnog udžbenika odobrio je Odbor za znanstveno-nastavnu literaturu Sveučilišta u Zagrebu rješenjem broj 02–608/1–1992. od 27. svibnja 1992. god.

CIP – Katalogizacija u publikaciji
Nacionalna i sveučilišna biblioteka, Zagreb

519.22(075.8)

PAUŠE, Željko
Uvod u matematičku statistiku / Željko Pauše. – Zagreb : Školska knjiga, 1993. – 405 str. : ilustr. ; 24 cm. – (Udžbenici Sveučilišta u Zagrebu = Manualia Universitatis studiorum Zagrabiensis)

Bibliografija: str. 399–401. – Kazalo.

930331034

Tisak: HRVATSKA TISKARA, Zagreb

DR. ŽELJKO PAUŠE

redovni profesor Građevinskog
fakulteta Sveučilišta u Zagrebu

UVOD U MATEMATIČKU STATISTIKU



ŠKOLSKA KNJIGA – ZAGREB 1993

Objavljivanje ovog djela novčano je pomoglo Ministarstvo znanosti, tehnologije i informatike Republike Hrvatske

Sadržaj

Predgovor

Poznato je da statistika i statističke metode imaju značajnu ulogu u suvremenim tokovima tehnološkog i društvenog razvitka. Planiranje i upravljanje u modernim visoko organiziranim društvima umnogome se zasniva na brojnim podacima o različitim pojavama i procesima karakterističnim za određene društvene zajednice.

Manipuliranje podacima, njihovo prikupljanje, sređivanje i analiziranje, zatim tumačenje i objašnjenje fenomena na koje se podaci odnose, te konačno praktična primjena rezultata analiza, zahtijevali su da se teorijski i znanstveno utemelje postupci i metode za rješavanje navedenih zadataka. Pojednostavnjeno može se reći da je upravo statistika ona znanstvena disciplina koja obuhvaća tu problematiku.

Povijesni počeci statistike povezani su s potrebama države za evidencijom stanovništva, vojske, proizvodnje i drugoga, a mogu se naći u vrlo dalekoj prošlosti (stara Kina, Perzija, Grčka i Rim). U XVII. stoljeću uvode se na njemačkim sveučilištima predavanja iz statistike, koja se odnose na problematiku državnih popisa. Razvitku ekonomske statistike značajnije su pridonijeli i engleski znanstvenici XVII. stoljeća (J. Grant, E. Halley i W. Petty), koji su istraživali zakonitosti u masovnim društvenim pojavama.

Teorijsku podlogu matematičke statistike počeli su razvijati već utemeljitelji teorije vjerojatnosti: J. Bernoulli (1654–1705), P. S. Laplace (1749–1782), S. D. Poisson (1781–1840) i C. F. Gauss (1777–1855). Najvažnije probleme matematičke statistike (procjena parametara, testiranje hipoteza i dr.) i glavne ideje za njihovo rješavanje postavili su čuveni statističari tzv. anglosaksonske statističke škole: F. Galton (1822–1911), K. Pearson (1857–1936), W. S. Goset (1876–1937), R. A. Fisher (1890–1962), J. Neyman (1894–1981) i drugi.

Matematička se statistika danas, kao i mnoge druge znanstvene discipline snažno razvija. U svijetu postoje brojni znanstvenici i institucije koji se bave statističkim istraživanjima. Također postoji opsežna literatura, a i nekoliko specijaliziranih znanstvenih i stručnih časopisa namijenjenih jedino statističkoj problematici.

Danas se statistika ili njezini pojedini dijelovi sve češće pojavljuju i kao nastavni predmeti, ili dijelovi nastavnih predmeta, na srednjim, a posebno na visokim školama i fakultetima. Uzrok je tome spoznaja da su određeni statistički sadržaji potrebni već i za opću naobrazbu, a pogotovo su neka znanja iz statistike nužna za uspješno usvajanje modernih spoznaja iz mnogih drugih nastavnih predmeta (fizika, biologija, psihologija, mnoge ekonomske i tehničke discipline i dr.).

Svrha je ove knjige da na postupan i sustavan način upozna čitatelja s najvažnijim pojmovima, načelima i metodama matematičke statistike, te da ga uputi u najvažnije primjene.

Pošlo se od pretpostavke da je najlakši i najprirodniji put da se to ostvari razmotriti najprije statističke fenomene koji se očituju uz pomoć statističkih podataka, zatim ukratko izložiti matematičku teoriju koja tumači i oblikuje

statističke zakonitosti i naposljetku prikazati metode i primjene teorije statističkog zaključivanja. Zbog toga je knjiga podijeljena na tri dijela: **Deskriptivnu statistiku, Matematičku teoriju statističkih fenomena i Teoriju statističkog zaključivanja.**

Prvi dio (Deskriptivna statistika) bavi se problemima manipuliranja statističkim podacima i njihova prikazivanja. Prikazane su različite metode tabličnog i grafičkog prikazivanja danog niza podataka, uvedeni su odgovarajući pojmovi za globalno opisivanje (razdioba frekvencija, grafikon frekvencija i sl.), a zatim su definirani odgovarajući parametri za opisivanje pojedinih karakteristika niza statističkih podataka (parametri lokacije, parametri rasipanja, parametri oblika itd.).

Navedeni sadržaji mogu se pratiti i usvajati na temelju predznanja elementarne matematike, tako da se tim dijelom knjige mogu služiti već i srednjoškolci. Osim toga, sadržaji deskriptivne statistike, koji su tijesno povezani s empirijskim fenomenima, važna su i nezaobilazna podloga za razumijevanje i lakše shvaćanje apstraktnih teorijskih pojmova teorije slučajnih varijabli i teorije statističkog zaključivanja. Čini se da bez uočavanja empirijskih fenomena, kao što su relativna frekvencija i razdioba frekvencija, i nije moguće shvatiti pravi smisao apstraktnih razdioba vjerojatnosti.

Iako je uobičajeno da se u knjige ovoga tipa uvrštavaju i elementi teorije vjerojatnosti, gdje se obično objašnjava pojam vjerojatnosti događaja i navode osnovne formule o vjerojatnosti, to ovdje nije učinjeno zbog dva razloga. Danas se, naime, već u mnogim srednjim školama obrađuju ti sadržaji, tako da se pretpostavlja da učenici završnih razreda tu građu znaju, pa se ispuštanjem te građe dobiva knjiga "čistije" koncepcije.

Drugi je razlog pretpostavka da je metodički pristup građi drugog dijela (Matematička teorija statističkih fenomena) tako načinjen da i nije nužno posebno razmatranje elemenata teorije vjerojatnosti. Naime, teorija slučajnih varijabli (diskretnih i kontinuiranih) pokušala se prikazati kao neposredna apstrakcija empirijskih statističkih obilježja razmotrenih u prvom dijelu knjige. U drugom dijelu stalno se naglašava da se teorijski pojmovi (razdioba vjerojatnosti, matematičko očekivanje, varijanca, momenti vjerojatnosne razdiobe i dr.) trebaju shvatiti i promatrati kao matematička apstrakcija "konkretnih" empirijskih pojmova (razdioba relativnih frekvencija, aritmetička sredina, varijanca, momenti i dr. niza statističkih podataka).

Za razliku od prvog i trećeg dijela knjige, gdje su se nastojali, osim definiranja i komentiranja uvedenih pojmova, iznijeti i jednostavniji dokazi, u drugom je dijelu namjerno izostavljeno dokazivanje svojstava i odnosa za uvedene pojmove, posebno ključnih teorema o funkcijama slučajnih varijabli (IV.6. i V.6) na kojima se uglavnom temelji teorija statističkog zaključivanja. To je učinjeno u prvom redu zbog toga što je za dokaze spomenutih teorema nužno vrlo opsežno i specijalno matematičko predznanje. Opća teorija slučajnih varijabli zapravo je dio teorije vjerojatnosti koja se, kako je poznato, aksiomatski zasniva i formalno-logički izgrađuje kao i sve druge aksiomatizirane matematičke teorije, pri čemu se primjenjuje moćan i sofisticiran matematički aparat (teorija mjere, opća teorija funkcija i dr.). Da bi se to izbjeglo u knjizi se promatraju samo diskretne i kontinuirane slučajne varijable i slučajni vektori, koji se mogu opisati relativno jednostavnim sredstvima matematičke analize i linearne algebre, a dovoljni su za razmatranje glavnih problema matematičke statistike.

Može se reći da je glavna namjena drugog dijela knjige da se na jasan i pregledan način iznesu i komentiraju najvažniji rezultati teorije slučajnih varijabli, nužnih za sustavno i logički konzistentno iznošenje teorije statističkog zaključivanja, kako bi se izbjeglo previše pozivanja na citiranu literaturu, koja čitatelju uvijek i nije dostupna.

Najvažniji i najopsežniji je treći dio knjige, koji obuhvaća prave sadržaje matematičke statistike, tako da obrazovaniji, u matematičkom smislu, čitatelj može preskočiti prva dva dijela knjige.

Pojednostavnjeno govoreći, središnje pitanje teorije statističkog zaključivanja jest što se može zaključiti o promatranoj pojavi na temelju određenoga konačnog niza podataka dobivenih mjerenjem (opažanjem) relevantnih veličina za tu pojavu. Tako postavljen problem očigledno je previše općenit i nejasan, tako da je potrebno usvojiti dodatne pretpostavke da bi se mogla razviti odgovarajuća teorija, koja će poslužiti kao oslonac za definiranje praktičnih postupaka statističkog zaključivanja.

Pri gruboj klasifikaciji može se reći da u teoriji statističkog zaključivanja postoje dvije grupe problema – procjena parametara i testiranje hipoteza. Procjeni parametara posvećeno je VI. (točkasta procjena) i VII. (intervalna procjena) poglavlje, testiranju hipoteza VIII, IX. i X. poglavlje, dok se u XI, XII, XIII. i XIV. poglavlju isprepleću obje problematike.

Budući da je teorija statističkog zaključivanja, zapravo, sastavljena od skupine matematičkih modela prilagođenih i namijenjenih rješavanju određenih praktičnih problema, nastojalo se gradivo iznijeti tako da se jasno razluči praktični aspekt problema od teorijskoga. Zbog onih čitatelja koje prije svega zanima praktični aspekt i neposredna primjena metoda statističkog zaključivanja, nastojala su se izbjeći duga i teška matematička izvođenja, pa se na nekim mjestima čitatelj upućuje na navedenu literaturu, a mnogi teorijski izvodi i dokazi prebačeni su u zadatke (uz uputu) na kraju odgovarajućeg poglavlja. Neki temeljni teoremi (Rao-Cramerova nejednakost, Neyman-Pearsonova lema, Gauss-Markovljev teorem i dr.) ipak su potpuno izvedeni, jer se to moglo načiniti vrlo jednostavnim sredstvima, a zahtjevnijeg čitatelja može potaknuti na dublje i studioznije upoznavanje matematičke statistike.

Statističko zaključivanje je suptilan i poseban način zaključivanja, koji je pojmovno vrlo težak i zahtijeva pojačanu koncentraciju i umni napor. Iako se sam postupak odlučivanja može prilično šablonizirati i svesti na rutinske operacije jednostavnog računanja s danim podacima te na primjenu odgovarajućih tablica, ili što je danas još češće na primjenu računala, shvaćanje pravog smisla i biti izvedenih zaključaka nije nimalo jednostavno.

Statističke metode imaju bitna ograničenja, tako da se ne smiju primjenjivati bez odgovarajućih spoznaja o njihovom stvarnom dometu i pravom značenju u svakoj konkretnoj situaciji. Teorijska naobrazba istraživača i stručnjaka koji se koriste statističkim metodama mogu biti jamstvo da će se one ispravno upotrebljavati, tako da se ovom knjigom želi pridonijeti izgradnji i podizanju statističke naobrazbe svih onih kojima je to potrebno.

Popis oznaka

$\in$	...pripada skupu...
$\notin$	...ne pripada skupu...
$\subseteq$	...je podskup od...
$\Rightarrow$	...implicira (povlači)...
$\to$	...konvergira (teži)...
$\mapsto$	...je pridruženo...
$\sim$	...ima razdiobu...
$\cup$	unija (skupova)
$\cap$	presjek (skupova)
$\setminus$	razlika (skupova)
$\times$	Kartezijev produkt (skupova)
$\forall$	za svaki
$\mathbf{N}$	skup svih prirodnih brojeva
$\mathbf{Z}$	skup svih cijelih brojeva
$\mathbf{R}$	skup svih realnih brojeva
$\mathbf{R}^n$	skup svih uređenih n-torki realnih brojeva
(a,b)	uređeni par a, b
$\langle a,b\rangle$	interval (otvoreni) a, b
$[a,b]$	segment (zatvoreni interval) a, b
$\lVert \cdot \rVert$	euklidska norma (vektora)
F^{-1}	inverzna funkcija od F
$\boldsymbol{A}^\top$	transponirana matrica od $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	inverzna matrica od $\boldsymbol{A}$
$\Gamma(\cdot)$	gama funkcija
$\mathrm{P}(\cdot)$	vjerojatnost (događaja)
$\mathrm{P}(\cdot/\cdot)$	uvjetna vjerojatnost
$\mathrm{E}[\cdot]$	matematičko očekivanje (slučajne varijable)
$\mathrm{V}[\cdot]$	varijanca (slučajne varijable)
$\mathrm{Cov}(\cdot,\cdot)$	kovarijanca (slučajnih varijabli)
$\boldsymbol{\Sigma}$	kovarijancna matrica (slučajnog vektora)
$\mathrm{B}(r,p)$	binomna razdioba s parametrima r i p
$\mathrm{B}(1,p)$	Bernoullijeva razdioba s parametrom p
$\mathrm{Po}(\lambda)$	Poissonova razdioba s parametrom λ
$\mathrm{N}(\mu,\sigma^2)$	normalna (Gaussova) razdioba s parametrima μ i σ^2
$\mathrm{G}(\alpha,\beta)$	gama-razdioba s parametrima α i β
$\mathrm{Ex}(\alpha)$	eksponencijalna razdioba s parametrom α
$\chi^2(n)$	hikvadrat razdioba s n stupnjeva slobode
$\mathrm{U}(a,b)$	jednolika (uniformna) razdioba nad segmentom $[a,b]$
$\mathrm{LN}(\mu,\sigma^2)$	lognormalna razdioba s parametrima μ i σ^2

$\mathrm{t}(n)$	Studentova t-razdioba sa n stupnjeva slobode
$\mathrm{F}(r,s)$	F-razdioba sa (r,s) stupnjeva slobode
$\mathrm{N}(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\varrho)$	dvodimenzionalna normalna razdioba
$\mathrm{B}(r,p_1,p_2)$	trinomna razdioba
$\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	višedimenzionalna normalna razdioba s vektorom očekivanja $\boldsymbol{\mu}$ i kovarijancnom matricom $\boldsymbol{\Sigma}$
Φ	f.r.v. za $\mathrm{N}(0,1)$
H_n	f.r.v. za $\chi^2(n)$
G_n	f.r.v. za $t(n)$
$F_{r,s}$	f.r.v. za $F(r,s)$
$\overline{x}$	aritmetička sredina, prosjek (niza podataka)
$\overline{X}$	uzoračka aritmetička sredina
$s_0^2,\ \widehat{\sigma}^2$	varijanca (niza podataka)
$\widehat{\Sigma}^2$	uzoračka varijanca
σ	standardna devijacija (slučajne varijable)
$s_0,\ \widehat{\sigma}$	standardna devijacija (niza podataka)
$\widehat{\Sigma}$	uzoračka standardna devijacija
s^2	korigirana varijanca (niza podataka)
S^2	korigirana uzoračka varijanca
M	medijan (slučajne varijable)
$m,\ \widehat{m}$	medijan (niza podataka)
$\widehat{M}$	uzorački medijan
K	koeficijent asimetrije (niza podataka)
κ	koeficijent asimetrije (slučajne varijable)
E	koeficijent spljoštenosti (niza podataka)
ε	koeficijent spljoštenosti (slučajne varijable)
t	teorijski parametar
Θ	skup dopuštenih vrijednosti parametra t
$\widehat{T}$	procjenitelj parametra t
$\widehat{t}$	vrijednost procjenitelja $\widehat{T}$
γ	pouzdanost intervala povjerenja
z_γ	$\Phi^{-1}\left(\frac{1+\gamma}{2}\right)$
τ_γ	$G_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)$
λ_γ	$(1-\gamma)^{-\frac{1}{2}}$
H_0	nul-hipoteza
α	razina značajnosti

Popis kratica

pogl.	poglavlje
sl.	slika
tabl.	tablica
zad.	zadatak
f.g.v.	funkcija gustoće vjerojatnosti
f.r.v.	funkcija razdiobe vjerojatnosti
s.v.	slučajna varijabla
s.vk.	slučajni vektor
CGT	centralni granični teorem
ML	maximum likelihood
MNK	metoda najmanjih kvadrata
LN	linearni nepristrani
NLN	najbolji linearni nepristrani
GM	Gauss-Markov
KS	Kolmogorov-Smirnov
ANOVA	analysis of variance
DW	Durbin-Watson
MWW	Mann-Whitney-Wilcoxon

PRVI DIO

DESKRIPTIVNA STATISTIKA

Gotovo da i nema istraživačke djelatnosti u kojoj se bar u nekoj fazi ne pojavljuje potreba za prikupljanjem i analiziranjem određenih podataka o istraživanoj pojavi. U prirodnim i tehničkim znanostima istraživanja su redovito povezana s mjerenjima određenih fizikalnih veličina, tako da se rezultati mjerenja izražavaju brojevima, pa se govori o *brojčanim* ili *numeričkim podacima.* U nekim drugim istraživanjima rezultati se mogu iskazati određenim kvalitativnim atributima (boja, oblik, politički stav i sl.).

Ako na rezultate mjerenja, odnosno opažanja, djeluju tzv. slučajni faktori, onda se govori o *statističkim podacima.* To znači da je priroda promatrane pojave takva da se ne mogu kontrolirati svi mogući utjecaji na proces koji dovodi do konačnog rezultata. Stoga se usvaja pretpostavka da izmjereni statistički podaci sadrže i odgovarajuću slučajnu komponentu.

U *deskriptivnoj statistici* razvijene su određene metode i postupci za egzaktno proučavanje statističkih podataka. Pod tim se razumjeva sređivanje, prikazivanje i interpretiranje statističkih podataka, definiranje glavnih parametara, utvrđivanje njihovih teorijskih svojstava i njihova praktičnog značenja.

I. Prikazivanje statističkih podataka

1. Tablica frekvencija

Prilikom opažanja ili eksperimentiranja pažnja istraživača redovito je usmjerena na jednu ili više veličina. Ako se promatra samo jedna veličina, označimo je sa X, onda je rezultat jednog mjerenja jedan realan broj x. Višestrukim ponavljanjem mjerenja veličine X dobiva se konačni niz brojeva $x_1, \ldots, x_n$, kao rezultat n ponovljenih mjerenja. Veličina X obično se naziva ***statističko obilježje***, a dobiveni niz brojeva $x_1, \ldots, x_n$ statistički podaci o promatranome statističkom obilježju X.

1. primjer

Neka X označuje ocjenu iz matematike izraženu jednim od brojeva 1,2,3,4 i 5. Da bi se uvidjeli rezultati nastave matematike u jednom razredu od $n = 30$ učenika, promotrit će se njihove ocjene iz matematike na kraju školske godine. Uvidom u "imenik" dobiven je ovaj niz statističkih podataka:

$$1, 4, 2, 3, 1, 1, 2, 4, 2, 3, 4, 5, 3, 2, 2, 2, 5, 3, 2, 2, 3, 3, 4, 2, 3, 2, 3, 3, 3, 2\,.$$

Vidi se da je $x_1 = 1$, $x_2 = 4$, $x_3 = 2$, $x_4 = 3$ itd. Odmah se uočava da se neki brojevi pojavljuju u danom nizu više puta. Tako se broj 1 pojavljuje 3 puta, broj 2 pojavljuje se 11 puta, broj 3 pojavljuje se 10 puta, 4 se pojavljuje 4 puta i 5 se pojavljuje 2 puta. Govori se još da ocjeni 1 pripada frekvencija 3, ocjeni 2 frekvencija 11, ocjeni 3 frekvencija 10, ocjeni 4 frekvencija 4 i ocjeni 5 frekvencija 2.

Ako statističko obilježje X poprima samo vrijednosti iz nekoga diskretnog (konačnog ili prebrojivog) skupa A, onda se kaže da je X *diskretno obilježje*. U tom se slučaju prilikom mjerenja (opažanja) kao rezultati dobivaju elementi skupa A, pa se za svaki $a \in A$ može uočiti broj f njegova pojavljivanja u nizu od n mjerenja (opažanja) obilježja X. Broj f ($f \in \{0, 1, 2, \ldots\}$) zove se ***frekvencija***, a broj $p = \dfrac{f}{n}$ ***relativna frekvencija*** vrijednosti a u nizu statističkih podataka $x_1, \ldots, x_n$.

U navedenom primjeru X (ocjena iz matematike) je diskretno statističko obilježje i $A = \{1, 2, 3, 4, 5\}$ je pripadni skup mogućih vrijednosti. Broju $1 \in A$ pripada frekvencija $f_1 = 3$ i relativna frekvencija $p_1 = \dfrac{3}{30} = 0{,}1$. Broju $2 \in A$ pripada frekvencija $f_2 = 11$ i relativna frekvencija $p_2 = \dfrac{11}{30} \approx 0{,}37$ itd.

Za pregledno prikazivanje statističkih podataka, uz primjenu pojma frekvencije i relativne frekvencije, najčešće se upotrebljava *tablica frekvencija*.

Tablica 1.

Vrijednost obilježja X	Frekvencija	Relativna frekvencija
a_1	f_1	p_1
a_2	f_2	p_2
$\vdots$	$\vdots$	$\vdots$
a_r	f_r	p_r

Brojevi $a_1, \dots, a_r$ elementi su skupa A i u tabl. 1. obično se redaju po veličini od manjih prema većima ($a_1 < a_2 < \dots < a_r$), dok su $f_1, \dots, f_r$ pripadne frekvencije, a $p_1, \dots, p_r$ odgovarajuće relativne frekvencije.

2. primjer

Za statističke podatke iz 1. primjera tablica frekvencija izgleda ovako:

Tablica 2.

Ocjena	Frekvencija	Relativna frekvencija
1	3	$\frac{1}{10} = 0{,}10$
2	11	$\frac{11}{30} \approx 0{,}37$
3	10	$\frac{1}{3} \approx 0{,}33$
4	4	$\frac{2}{15} \approx 0{,}13$
5	2	$\frac{1}{15} \approx 0{,}07$
	zbroj = 30	zbroj = 1

Zbroj svih frekvencija iznosi $n = 30$, tj. jednak je broju izvršenih mjerenja (opažanja), dok je zbroj svih relativnih frekvencija 1.

Općenito vrijedi

(1) $$f_1 + f_2 + \dots + f_r = \sum_{j=1}^{r} f_j = n\,,$$

(2) $$0 \le p_j \le 1\,, \qquad j = 1, 2, \dots, r\,,$$

(3) $$p_1 + p_2 + \dots + p_r = \sum_{j=1}^{r} p_j = 1\,.$$

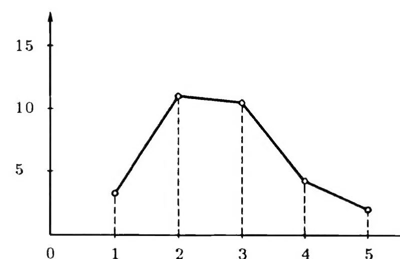
Relacija (1) neposredno slijedi iz definicije frekvencije. Budući da je $0 \le f_j \le n$, onda je $0 \le \frac{f_j}{n} \le 1$, a to se upravo izriče relacijom (2). Uvrštavanjem $p_j = \frac{1}{n} f_j$ u $\sum_{j=1}^{r} p_j$, dobiva se

$$\sum_{j=1}^{r} p_j = \sum_{j=1}^{r} \frac{f_j}{n} = \frac{1}{n} \sum_{j=1}^{r} f_j = \frac{1}{n} \cdot n = 1\,,$$

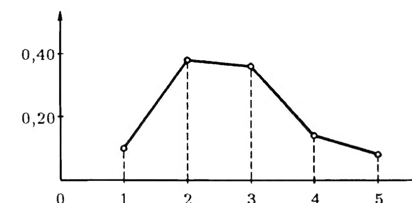
a to se upravo tvrdi u (3).

2. Grafikon frekvencija

Na temelju tabličnog prikaza statističkih podataka izrađuju se različiti grafički prikazi. Ako se na apscisnu os pravokutnoga koordinatnog sustava u ravnini nanesu vrijednosti (podaci) obilježja X, a kao pripadne ordinate uzmu odgovarajuće frekvencije (relativne frekvencije), dobiva se *grafikon frekvencija* (relativnih frekvencija) danog niza statističkih podataka. Spajanjem tako dobivenih točaka dobiva se odgovarajući *poligon frekvencija*. Očigledno je da između grafikona na sl. 1. i sl. 2. nema bitnih razlika, jer je zapravo riječ samo o promjeni mjerila na ordinatnoj osi.



Slika 1. Poligon frekvencija za podatke iz tabl. 2.



Slika 2. Poligon relativnih frekvencija za podatke iz tabl. 2.

Govori se da je tablicom frekvencija, odnosno grafikonom frekvencija (relativnih frekvencija), zadana ***razdioba frekvencija*** u danom nizu statističkih podataka. Funkcija $f : A \to \mathbf{R}$, definirana formulom

(4) $$f(a_j) = f_j\,, \qquad a_j \in A\,, \qquad j = 1, \dots, r\,,$$

zove se *funkcija frekvencija*, a funkcija $p : A \to \mathbf{R}$, definirana formulom

(5) $$p(a_j) = p_j\,, \qquad a_j \in A\,, \qquad j = 1, \dots, r\,,$$

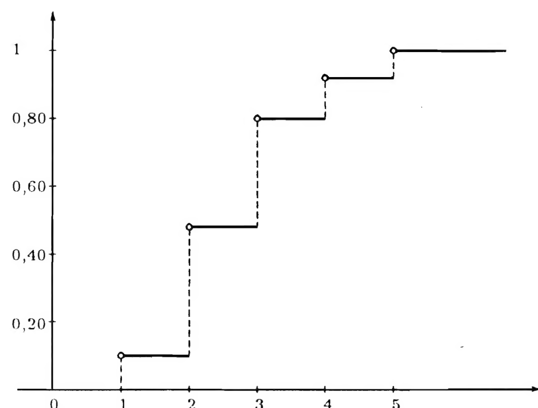
zove se *funkcija relativnih frekvencija* danog niza statističkih podataka o diskretnome statističkom obilježju X.

Uobičajeno je da se definira i tzv. *funkcija kumulativnih frekvencija*

(6a) $$K(x) = \sum_{a_j \le x} f_j\,, \qquad x \in \mathbf{R}\,,$$

odnosno *funkcija kumulativnih relativnih frekvencija*

(6b) $$F(x) = \sum_{a_j \le x} p_j\,, \qquad x \in \mathbf{R}\,.$$



Slika 3. Graf funkcije kumulativnih relativnih frekvencija za podatke iz tabl. 2

Za ilustraciju definiranih pojmova razmotrimo još jedan primjer.

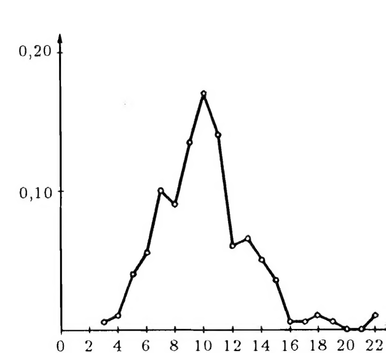
3. primjer

U industrijskom pogonu gdje se radi na velikom broju istovrsnih strojeva bilježen je dnevni broj kvarova na tim strojevima. Izvršeno je $n = 200$ opažanja, pri čemu su dobiveni podaci o statističkom obilježju X koje označuje dnevni broj kvarova. Podaci su odmah sređeni tako da je načinjena tabl. 3.

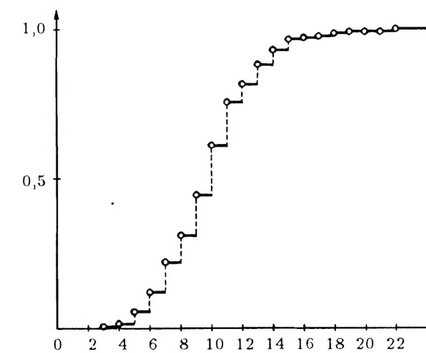
U ovom je primjeru $r = 20$, $a_1 = 3$, $a_2 = 4, \ldots a_{20} = 22$ i u tabl. 3. su navedene pripadne frekvencije, relativne frekvencije, kumulativne frekvencije i kumulativne relativne frekvencije. Na temelju tabl. 3. odmah se može načiniti odgovarajući grafikon relativnih frekvencija (sl. 4). Iz sl. 4. zorno se razabire da velike frekvencije pripadaju brojevima 9, 10 i 11, a također i ostala svojstva razdiobe frekvencija u danom skupu statističkih podataka. Tablica 3. omogućuje da se nacrta i pripadni graf funkcije kumulativnih relativnih frekvencija (sl. 5).

Tablica 3.

Redni broj	Broj kvarova	Frekvencija	Relativna frekvencija	Kumulativna frekvencija	Kumulativna relativna frekvencija
1	3	1	0,005	1	0,005
2	4	2	0,010	3	0,015
3	5	8	0,040	11	0,055
4	6	13	0,065	24	0,120
5	7	20	0,100	44	0,220
6	8	18	0,090	62	0,310
7	9	27	0,135	89	0,445
8	10	34	0,170	123	0,615
9	11	28	0,140	151	0,755
10	12	12	0,060	163	0,815
11	13	13	0,065	176	0,880
12	14	10	0,050	186	0,930
13	15	7	0,035	193	0,965
14	16	1	0,005	194	0,970
15	17	1	0,005	195	0,975
16	18	2	0,010	197	0,985
17	19	1	0,005	198	0,990
18	20	0	0,000	198	0,990
19	21	0	0,000	198	0,990
20	22	2	0,010	200	1,000



Slika 4. Grafikon relativnih frekvencija za podatke iz tabl. 3.



Slika 5. Grafikon funkcije kumulativnih relativnih frekvencija za podatke iz tabl. 3.

3. Prikazivanje podataka nenumeričkoga statističkog obilježja

Ako statističko obilježje nije numeričko, onda neki od definiranih pojmova nemaju smisla, jer skup A nije brojčani skup.

4. primjer

Svaki napisani tekst može se shvatiti kao niz statističkih podataka o slovima abecede. Na temelju toga niza također se može načiniti pripadna tablica frekvencija. Pogledajmo kako izgleda tablica frekvencija slova abecede, ako se kao niz podataka

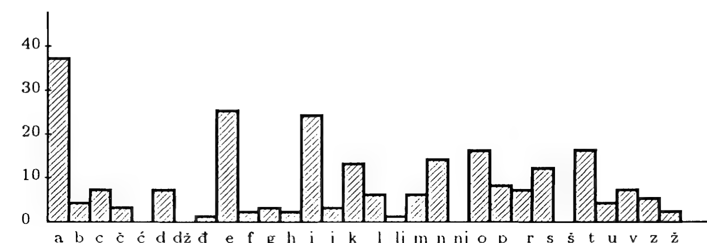
Tablica 4.

Redni broj	Slovo	Frekvencija	Relativna frekvencija
1	a	37	0,157
2	b	4	0,017
3	c	7	0,030
4	č	3	0,012
5	ć	0	0
6	d	7	0,030
7	dž	0	0
8	đ	1	0,004
9	e	25	0,016
10	f	2	0,008
11	g	3	0,012
12	h	2	0,008
13	i	24	0,102
14	j	3	0,012
15	k	13	0,055
16	l	6	0,026
17	lj	1	0,004
18	m	6	0,026
19	n	14	0,060
20	nj	0	0
21	o	16	0,068
22	p	8	0,034
23	r	7	0,030
24	s	12	0,050
25	š	0	0
26	t	16	0,068
27	u	4	0,017
28	v	7	0,030
29	z	5	0,020
30	ž	2	0,008

uzmu upravo napisane tri rečenice ($n = 235$). Pri brojenju se ne razlikuju velika i mala slova.

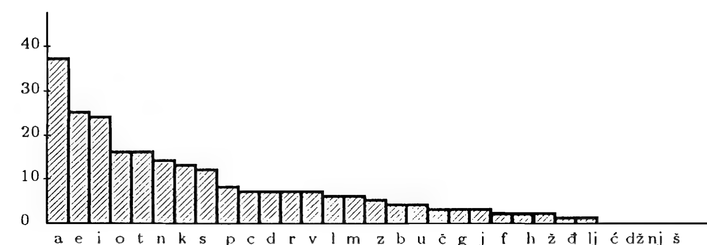
Očigledno je da se sada ne može govoriti o grafikonu frekvencija i poligonu frekvencija u onom smislu kako je to bilo kod brojčanoga diskretnog statističkog obilježja. Skup $A = \{a, b, c, \ldots, v, z, ž\}$ vrijednosti statističkog obilježja ima kao svoje elemente slova i stoga se ne može formirati koordinatni sustav za prikazivanje razdiobe frekvencija.

Da bi se ipak i zorno geometrijski uočila razdioba frekvencija po slovima abecede, može se nacrtati *histogram frekvencija* (relativnih frekvencija) tako da se iznad ispisanog slova nacrta pravokutnik visine jednake frekvenciji (relativnoj frekvenciji) toga slova u danom nizu (tekstu).



Slika 6. Histogram frekvencija za podatke iz 4. primjera

Izgled histograma frekvencija ovisi, dakako, o primijenjenom redoslijedu slova. Na sl. 6. primijenjen je tzv. abecedni redoslijed, međutim može se primijeniti i neki drugi redoslijed. Načini li se redoslijed slova tako da na prvo mjesto dođe najfrekventnije slovo, a zatim se po opadajućim frekvencijama poredaju ostala slova, pripadni histogram frekvencija izgleda posve drukčije (sl. 7).



Slika 7. Histogram po opadajućim frekvencijama za podatke iz 4. primjera

Za određena lingvistička istraživanja veću vrijednost ima histogram frekvencija na sl. 7. od onoga na sl. 6, jer se neke lingvističke zakonitosti, očigledno, lakše mogu uočiti na sl. 7. nego na sl. 6.

4. Kontinuirano statističko obilježje

Ako je riječ o statističkom obilježju X koje može poprimati vrijednosti iz nekog intervala skupa $\mathbf{R}$ realnih brojeva, onda se govori o *kontinuiranome statističkom obilježju* X.

5. primjer

Mjerenjem tlačne čvrstoće X tako da se načini $n = 100$ jednakih betonskih kocki, koje se izrađuju na odgovarajući način i zatim stavljaju u prešu i pritom se mjeri sila pri kojoj se kocka lomi, dobiveni su ovi rezultati (u MPa – megapaskalima):

30,97 , 42,63 , 35,76 , 45,00 , 40,15 , 38,79 , 47,12 , 33,56 , 39,22 , 34,47 , 32,54 ,
42,13 , 37,63 , 41,55 , 46,93 , 42,00 , 33,75 , 33,06 , 33,70 , 35,69 , 40,09 , 43,41 ,
40,32 , 35,73 , 36,87 , 30,16 , 40,83 , 36,65 , 30,96 , 36,36 , 32,72 , 36,73 , 29,70 ,
40,08 , 35,98 , 35,83 , 39,49 , 33,16 , 32,93 , 30,32 , 37,05 , 32,60 , 32,79 , 41,17 ,
39,48 , 37,54 , 33,09 , 40,74 , 34,82 , 37,52 , 37,93 , 30,09 , 43,36 , 36,17 , 35,59 ,
25,67 , 32,99 , 36,90 , 38,10 , 36,66 , 28,86 , 32,88 , 45,02 , 35,17 , 35,09 , 33,89 ,
27,26 , 35,94 , 33,77 , 29,50 , 27,10 , 36,39 , 32,22 , 38,89 , 30,81 , 37,60 , 30,18 ,
39,76 , 37,69 , 27,28 , 37,94 , 32,15 , 34,88 , 25,13 , 30,05 , 36,11 , 32,10 , 38,18 ,
36,69 , 33,01 , 29,18 , 40,82 , 39,01 , 33,86 , 32,79 , 31,56 , 28,36 , 37,06 , 33,09 ,
29,60 .

Dobiven je, dakle, niz od $n = 100$ statističkih podataka, pri čemu je $x_1 = 30,97$, $x_2 = 42,63$, $\ldots$, $x_{99} = 33,09$, $x_{100} = 29,60$.

Odmah se može uočiti da u ovom nizu od 100 podataka nema međusobno jednakih brojeva, što je donekle i razumljivo ako se ima na umu priroda mjerene veličine. U ovom je, naime, primjeru statističko obilježje X fizikalna veličina koja se mjeri odgovarajućom mjernom jedinicom (paskal) i teorijski gledano može poprimiti bilo koju vrijednost iz intervala $[0, \infty\rangle$. Budući da se raspolaže s vrlo velikim brojem mjerenja ($n = 100$), pri čemu se pojavljuje i velika raznolikost u rezultatima mjerenja, prirodno se nameće ideja da se izvrši određeno *grupiranje podataka* kako bi se dobila pregledna tablica frekvencija i odgovarajući grafički prikazi.

Da bi se to postiglo najprije se uoči najmanja ($x_{\min} = x_{84} = 25{,}13$) i najveća ($x_{\max} = x_7 = 47{,}12$) vrijednost u danom nizu statističkih podataka. Prema tome, svi se podaci nalaze u intervalu [25,13 ; 47,12]. Uzme li se taj interval, ili radi praktičnosti nešto širi interval I = [25,49], i razbije na 20 podintervala ili *razreda* širine 1,2 i zatim odredi frekvencija svakog razreda, tj. broj onih rezultata u danom nizu statističkih podataka koji pripadaju dotičnom razredu, dobiva se tablica frekvencija za grupirane podatke promatranoga statističkog obilježja X (tabl. 5).

Iz tabl. 5. vidi se da je 1. razred, zapravo, interval [25,0 ; 26,2⟩, 2. razred je interval [26,2 ; 27,4⟩ itd., sve do 20. razreda koji je, zapravo, interval [47,8 ; 49,0]. U tabl. 5. navedena je i sredina svakog razreda, te pripadne frekvencije, relativne frekvencije, i kumulativne frekvencije za dani niz statističkih podataka.

Tablica 5.

Redni broj razreda	Donji rub razreda	Gornji rub razreda	Sredina razreda	Frekvencija razreda	Relativna frekvencija razreda	Kumulativna relativna frekvencija
1	25,0	26,2	25,6	2	0,02	0,02
2	26,2	27,4	26,8	3	0,03	0,05
3	27,4	28,6	28,0	1	0,01	0,06
4	28,6	29,8	29,2	5	0,05	0,11
5	29,8	31,0	30,4	8	0,08	0,19
6	31,0	32,2	31,6	3	0,03	0,22
7	32,2	33,4	32,8	14	0,14	0,36
8	33,4	34,6	34,0	7	0,07	0,43
9	34,6	35,8	35,2	8	0,08	0,51
10	35,8	37,0	36,4	13	0,13	0,64
11	37,0	38,2	37,6	11	0,11	0,75
12	38,2	39,4	38,8	4	0,04	0,79
13	39,4	40,6	40,0	7	0,07	0,86
14	40,6	41,8	41,2	5	0,05	0,91
15	41,8	43,0	42,4	3	0,03	0,94
16	43,0	44,2	43,6	2	0,02	0,96
17	44,2	45,4	44,8	2	0,02	0,98
18	45,4	46,6	46,0	0	0,00	0,98
19	46,6	47,8	47,2	2	0,02	1,00
20	47,8	49,0	48,4	0	0,00	1,00

5. Grupiranje podataka u razrede

Primjer 5. upućuje nas kako da se općenito za dani niz $x_1, \ldots, x_n$ statističkih podataka o nekome kontinuiranom obilježju X načini pregledni prikaz pomoću odgovarajuće tablice frekvencija. Prvi korak sastoji se u definiranju razreda. Razredi su određeni disjunktni podintervali intervala svih mogućih vrijednosti kontinuiranoga statističkog obilježja X. Širine pojedinih razreda, u načelu, su proizvoljne i njihov izbor nije uvjetovan teorijskim razlozima, već praktičnim potrebama da tablica bude pregledna i da se mogu uočiti bitna svojstva promatranoga statističkog obilježja. Redovito se radi o razredima jednake širine i broj r razreda bira se obično u ovisnosti o broju n podataka u danom nizu. Postoje određene preporuke o izboru broja r u ovisnosti o broju n (r treba biti 5 – 10 % od n, ali ne veći od 30), ali to se zasniva na empirijskim spoznajama i nema teorijskog utemeljenja.

Tablični prikaz statističkih podataka o nekom kontinuiranom obilježju X, primjenom grupiranja podataka, općenito izgleda ovako:

Tablica 6.

Redni broj razreda	Donji rub razreda	Gornji rub razreda	Sredina razreda	Frekvencija razreda	Relativna frekvencija razreda	Kumulativna relativna frekvencija
1	a_0	a_1	$\overline{a}_1$	f_1	p_1	F_1
2	a_1	a_2	$\overline{a}_2$	f_2	p_2	F_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	a_{r-1}	a_r	$\overline{a}_r$	f_r	p_r	F_r

Broj a_0 izabran je tako da vrijedi $a_0 \leq x_{\min}$, a broj a_r tako da vrijedi $a_r \geq x_{\max}$. Nadalje vrijedi

$$a_0 < a_1 < a_2 < \ldots < a_{r-1} < a_r \,, \tag{7}$$

$$\overline{a}_j = \frac{1}{2}(a_{j-1} + a_j)\,, \qquad p_j = \frac{1}{n} f_j \qquad F_j = \sum_{i=1}^{j} p_i \,, \qquad j = 1, \ldots, r\,. \tag{8}$$

Ako je riječ o razredima jednake širine d, onda je

$$d = \frac{a_r - a_0}{r}\,, \qquad a_j = a_0 + jd\,, \qquad j = 1, \ldots, r\,. \tag{9}$$

Očigledno je da i u ovom slučaju vrijedi

$$\sum_{j=1}^{r} f_j = n\,, \qquad 0 \leq p_j \leq 1\,, \qquad \sum_{j=1}^{r} p_j = 1\,. \tag{10}$$

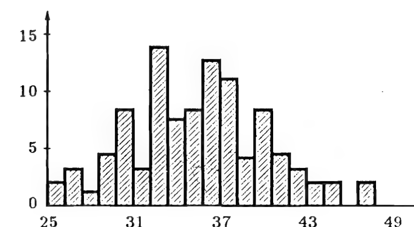
U 5. primjeru imali smo $r = 20$, $a_0 = 25$, $a_{20} = 49$ i $d = \frac{24}{20} = 1{,}2$, iz čega su proizašle vrijednosti $a_1 = 26{,}2$, $a_2 = 27{,}4$, $a_3 = 28{,}6$ itd. Iz tabl. 5. vidi se da je $f_1 = 2$, što znači da u danom nizu statističkih podataka postoje dva broja koja pripadaju razredu (intervalu) $[25\,;\,26{,}2\rangle$. To su brojevi $x_{84} = 25{,}13$ i $x_{56} = 25{,}67$. Također se vidi da je $f_2 = 3$, tj. da 2. razredu pripada frekvencija 3, što znači da postoje tri broja u danom nizu statističkih podataka koja se nalaze u intervalu $[26{,}2\,;\,27{,}4\rangle$. To su brojevi $x_{67} = 27{,}26$, $x_{71} = 27{,}10$ i $x_{80} = 27{,}28$.

Općenito u j-ti razred $[a_{j-1}, a_j\rangle$ $(j = 1, \ldots, r)$ ulaze svi oni podaci danog niza statističkih podataka koji su veći (ili jednaki) od donjeg ruba a_{j-1}, a manji od gornjeg ruba a_j toga razreda. Jedino r-ti razred sadrži i podatke koji su jednaki gornjem rubu a_r toga razreda.

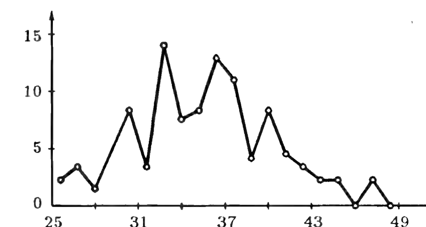
6. Histogram i poligon frekvencija

Na temelju tabličnog prikaza statističkih podataka načinjenog primjenom grupiranja podataka (tabl. 6) može se načiniti i odgovarajući grafički prikaz. Ako se na apscisnu os pravokutnoga koordinatnog sustava u ravnini nanesu vrijednosti rubova (granica) razreda i zatim iznad svakog razreda ucrta pravokutnik visine jednake odgovarajućoj frekvenciji (relativnoj frekvenciji) toga razreda, dobiva se

tzv. *histogram frekvencija* danog niza statističkih podataka koji, dakako, ovisi o izvedenom grupiranju podataka.



Slika 8. Histogram frekvencija za podatke iz tabl. 5.



Slika 9. Poligon frekvencija za podatke iz tabl. 5.

Ako se, pak, na apscisnu os nanesu sredine razreda, a kao pripadne ordinate uzmu odgovarajuće frekvencije (relativne frekvencije), dobiva se poligon frekvencija danog niza statističkih podataka uz provedeno grupiranje podataka u razrede.

Poligon frekvencija može se interpretirati i kao linearna interpolacija funkcije definirane formulom

$$f(\overline{a}_j) = f_j\,, \qquad j = 1, \ldots, r\,, \tag{11}$$

gdje je $\overline{a}_j$ sredina j-tog razreda. Ova funkcija također se zove *funkcija frekvencija*, a funkcija definirana formulom

$$p(\overline{a}_j) = p_j\,, \qquad j = 1, \ldots, r\,. \tag{12}$$

zove se *funkcija relativnih frekvencija* danog niza statističkih podataka o kontinuiranom obilježju X i primijenjenog grupiranja podataka.

Ovo razmatranje upućuje na činjenicu da je pri grupiranju podataka u razrede, zapravo, riječ o tome da se sve vrijednosti iz j-tog razreda aproksimiraju sredinom $\overline{a}_j$ toga razreda. Time se, dakako, gubi određeni dio informacije o promatranoj pojavi sadržane u izmjerenom nizu statističkih podataka, ali se, na drugoj strani, dobiva mogućnost da se razluče bitna svojstva promatranoga kontinuiranog statističkog obilježja X od nebitnih.

Kakve se promjene zbivaju kada se promijeni broj razreda, što povlači i promjenu širine razreda, ilustrirat će se idućim primjerom.

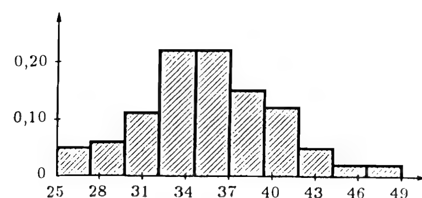
6. primjer

Ako se statistički podaci iz 5. primjera grupiraju u $r = 10$ razreda, dobiva se tablica frekvencija prikazana tablicom 7.

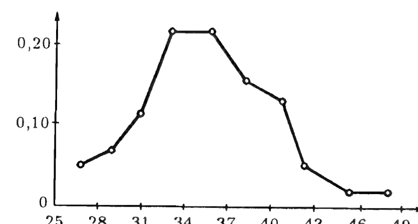
Sada je širina razreda $d = \frac{1}{10}(a_{10} - a_0) = 2{,}4$ i očigledno je da sredina razreda grublje aproksimira vrijednosti razreda nego pri grupiranju istih podataka u 20 razreda širine 1,2. No pogleda li se pripadni histogram relativnih frekvencija (sl. 10) za tabl. 7. i pripadni poligon relativnih frekvencija (sl. 11), očigledno je da se na njima lakše uočavaju određne pravilnosti u razdiobi frekvencija pri promatranju uočenoga statističkog obilježja X, nego što se to može vidjeti na sl. 8. i sl. 9.

Tablica 7.

Redni broj razreda	Donji rub razreda	Gornji rub razreda	Sredina razreda	Frekvencija razreda	Relativna frekvencija razreda	Kumulativna relativna frekvencija
1	25,0	27,4	26,2	5	0,05	0,05
2	27,4	29,8	28,6	6	0,06	0,11
3	29,8	32,2	31,0	11	0,11	0,22
4	32,2	34,6	33,4	21	0,21	0,43
5	34,6	37,0	35,8	21	0,21	0,64
6	37,0	39,4	38,2	15	0,15	0,79
7	39,4	41,8	40,6	12	0,12	0,91
8	41,8	44,2	42,4	5	0,05	0,96
9	44,2	46,6	45,4	2	0,02	0,98
10	46,6	49,0	47,8	2	0,02	1,00



Slika 10. Histogram relativnih frekvencija za podatke iz tabl. 7.



Slika 11. Poligon relativnih frekvencija za podatke iz tabl. 7.

7. Funkcija kumulativnih frekvencija

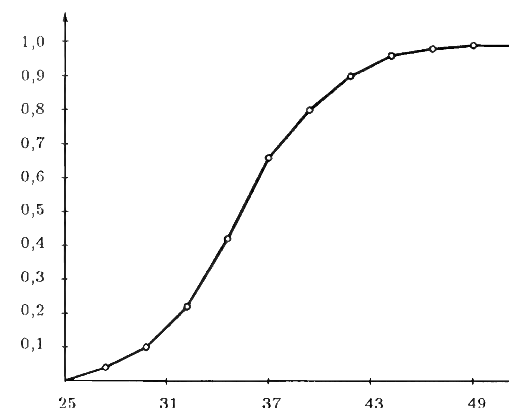
Za podatke iz tabl. 7. može se nacrtati i graf funkcije $x \mapsto F(x)$, $x \in \mathbf{R}$, kumulativnih relativnih frekvencija, što je prikazano na sl. 12.

Krivulja na sl. 12. nacrtana je tako da su najprije dužinom spojene točke (25,0 ; 0) i (27,4 ; 0,05), zatim su dužinom spojene točke (27,4 ; 0,05) i (29,8 ; 0,11), pa točke (29,8 ; 0,11) i (32,2 ; 0,22) itd.

Općenito se graf funkcije kumulativnih relativnih frekvencija $x \mapsto F(x)$ za podatke o kontinuiranome statističkom obilježju crta tako da se dužinama spoje točke s koordinatama (a_j, F_j) i (a_{j+1}, F_{j+1}) $(j = 1, \ldots, r)$, pri čemu je početna točka $(a_0, 0)$. Očigledno je da se može pisati

$$F_j = F(a_j) \qquad j = 1, \ldots, r .$$

Time se dobiva krivulja, poput one na sl. 12, koja se sastoji od dužina. Na lijevom kraju krivulja počinje na apscisnoj osi (točka $(a_0, 0)$), a na desnom kraju završava u točki $(a_r, 1)$. Unutar intervala $[a_0, a_r]$ krivulja je monotono rastuća i zorno pokazuje (svojim nagibom) kako se nakupljaju (kumuliraju) dani statistički



Slika 12. Graf funkcije kumulativnih relativnih frekvencija za podatke iz tabl. 7.

podaci duž apscisne osi. Strmiji nagib krivulje odgovara većoj brzini kumuliranja podataka.

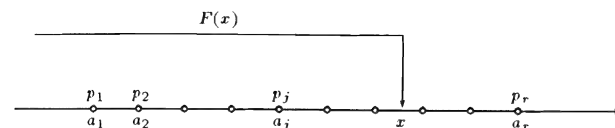
Krivulja kumulativnih relativnih frekvencija produžuje se i izvan intervala $[a_0, a_r]$, i to ulijevo po osi apscisa, a udesno po pravcu usporednom s apscisnom osi na visini jedan. Izvan intervala $[a_0, a_r]$ nagib krivulje je nula, a to znači da u tom dijelu apscisne osi nema brojeva iz danog niza statističkih podataka. Može se dogoditi da krivulja kumulativnih relativnih frekvencija bude usporedna s apscisnom osi i na nekom podintervalu intervala $[a_0, a_r]$ i to onda znači da u tom području nema brojeva iz promatranog niza statističkih podataka.

Sada se, slično kao i kod diskretnoga statističkog obilježja, može reći da je tablicom frekvencija te pripadnim histogramom frekvencija, odnosno krivuljom kumulativnih relativnih frekvencija, definirana razdioba frekvencija u danom nizu statističkih podataka o promatranome kontinuiranom statističkom obilježju X. Važno je, ipak, primijetiti da postoje bitne razlike između razdiobe frekvencija kod diskretnog i kod kontinuiranog obilježja. Kod diskretnog obilježja frekvencije (relativne frekvencije) se pridružuju brojevima, dok se kod kontinuiranog obilježja frekvencije pridružuju razredima, tj. brojevnim intervalima. Stoga danim nizom statističkih podataka o nekom kontinuiranom obilježju i nije jednoznačno određena pripadna razdioba frekvencija, što je slučaj kod diskretnog obilježja. Razdioba frekvencija kod kontinuiranoga statističkog obilježja ovisi o načinu grupiranja podataka u razrede.

8. Mehanička interpretacija razdiobe frekvencija

Zanimljivo je primijetiti da postoji određena pojmovna i računska analogija između pojma "razdioba frekvencija (relativnih frekvencija)" i mehaničkog pojma "linijska razdioba (distribucija) mase".

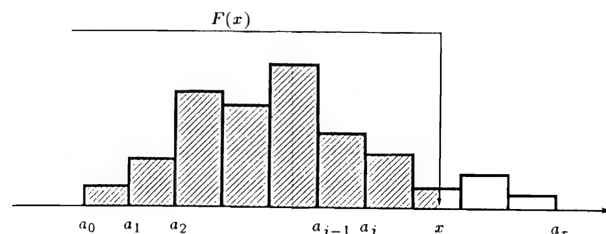
Ako se pri diskretnome statističkom obilježju X brojevi $a_1, \ldots, a_r$ (v. tabl. 1) interpretiraju kao apscise točaka na brojevnoj osi, a brojevi $f_1, \ldots, f_r$ (frekvencije), odnosno brojevi $p_1, \ldots, p_r$ (relativne frekvencije), kao pripadne mase, onda se može govoriti o određenoj razdiobi ukupne mase n, odnosno 1, po danom pravcu. Prema tome, može se tumačiti da je tablicom 1. definirana određena linijska razdioba mase, pa se u tom smislu mogu tumačiti i formule (1), (2), (3) te (6a i 6b). Formula (6a), na primjer, pokazuje onu količinu mase koja je raspodijeljena lijevo od točke s apscisom x ($x \in \mathbf{R}$), uključujući i tu točku.



Slika 13. Linijska razdioba mase diskretnog tipa

Analogija između razdiobe frekvencija i linijske razdiobe mase ima dalekosežne posljedice i omogućuje lakše usvajanje i zorno predstavljanje mnogih pojmova koji će se definirati u vezi sa statističkim podacima.

I u slučaju kontinuiranoga statističkog obilježja može se uspostaviti analogija između razdiobe frekvencija i linijske razdiobe mase. Tada se, naime, frekvencija f_j, odnosno relativna frekvencija razreda p_j, interpretira kao jednoliko kontinuirano raspodijeljena masa na intervalu $[a_{j-1}, a_j\rangle$ brojevne osi. Stoga se može tumačiti da je tablicom 6. definirana određena kontinuirana linijska razdioba ukupne mase n, odnosno 1, po danom pravcu.



Slika 14. Kontinuirana linijska razdioba mase

Veličina $F(x)$ pokazuje onu količinu mase koja je raspodijeljena lijevo od točke $x (x \in \mathbf{R})$.

Primjedba

Za sređivanje i prikazivanje statističkih podataka danas se uvelike upotrebljavaju računski strojevi, napose tzv. osobna računala (personal computer, PC) i dodatna oprema koja ide uz njih (monitori, pisači i sl.). Da bi se što više olakšala primjena računala u obradi statističkih podataka, izrađeni su tzv. *programski paketi* za pojedine tipove zadataka. Nakon unošenja u računalo, opskrbljeno odgovarajućim programskim paketom, niza statističkih podataka, jednostavnim postupkom omogućeno je dobivanje odgovarajućih tablica frekvencija i relativnih frekvencija, te pripadnih grafikona, histograma i drugih oblika grafičkog prikaza danih podataka.

Zadaci

Za statističke podatke o diskretnom statističkom obilježju X u zadacima 1–7. načinite pripadnu tablicu frekvencija i relativnih frekvencija, grafikon frekvencija i poligon relativnih frekvencija, te graf funkcije kumulativnih relativnih frekvencija.

1. X – broj dobiven bacanjem igraće kocke: 1, 3, 1, 6, 2, 6, 4, 6, 3, 3, 4, 3, 1, 4, 4, 1, 4, 5, 3, 5, 4, 1, 1, 1, 5, 4, 3, 2, 1, 4, 6, 4, 3, 2, 2, 2, 3, 6, 1, 2, 2, 3, 5, 3, 1, 3, 6, 2, 4, 1.
2. X – broj odsutnih učenika na satu matematike: 2, 5, 1, 1, 3, 4, 4, 4, 2, 3, 3, 4, 0, 0, 4, 4, 3, 6, 1, 4, 2, 2, 4, 3, 2, 1, 3, 2, 2, 5, 4, 0, 3, 2, 1, 2, 4, 1, 3, 3.
3. X – dnevni broj prometnih nezgoda na određenoj cesti: 0, 0, 2, 3, 2, 1, 4, 2, 2, 1, 2, 1, 4, 0, 2, 1, 2, 0, 1, 1, 1, 2, 1, 0, 4, 2, 2, 2, 1, 1, 1, 1, 0, 2, 2, 0, 3, 2, 3, 4.
4. X – broj dana bez oborina u mjesecu rujnu: 21, 16, 20, 17, 18, 21, 13, 18, 22, 15, 19, 19, 16, 14, 22, 16, 17, 16, 14, 16.
5. X – dnevni broj prodanih pari cipela u nekoj prodavaonici obuće: 99, 103, 94, 97, 89, 101, 105, 87, 95, 124, 105, 100, 107, 90, 114, 119, 99, 105, 84, 96, 103, 94, 81, 96, 103, 112, 85, 96, 98, 110, 99, 113, 112, 96, 115, 114, 100, 104, 104, 94.
6. X – tjedni broj kvarova na strojevima nekoga industrijskog pogona: 0, 0, 1, 1, 0, 3, 1, 1, 0, 1, 1, 2, 2, 0, 3, 2, 2, 0, 0, 1, 4, 1, 0, 2, 3, 2, 1, 1, 3, 2, 1, 1, 2, 0, 1, 3, 2, 1, 1, 2, 2, 3, 2, 1, 1, 2, 0, 1, 2, 1.
7. X – broj telefonskih razgovora preko određene telefonske centrale u jednom satu: 24, 35, 29, 28, 29, 16, 31, 26, 23, 33, 25, 19, 32, 31, 23, 16, 17, 26, 23, 27, 27, 24, 22, 23, 13.

Statističke podatke o kontinuiranome statističkom obilježju X u zadacima 8 – 14. grupirajte u razrede, načinite pripadne tablice frekvencija i relativnih frekvencija, nacrtajte histogram i poligon frekvencija, te graf funkcije kumulativnih relativnih frekvencija.

8. X – vlačna čvrstoća čelične žice (MPa): 285, 341, 323, 300, 313, 294, 305, 317, 286, 312, 267, 316, 300, 298, 312, 319, 296, 284, 293, 298, 322, 292, 267, 305, 299, 275, 318, 304, 298, 301, 282, 309, 297, 313, 296, 323, 305, 307, 289, 307, 396, 342, 310, 336, 286, 320, 290, 323, 288, 306.

9. X – vijek trajanja žarulje (sati): 0,4 , 53,3 , 254,7 , 41,1 , 220,6 , 201,6 , 73,4 , 143,3 , 108,8 , 131,7 , 54,0 , 233,1 , 29,1 , 17,8 , 13,9 , 143,5 , 520,8 , 318,4 , 45,9 , 246,4 , 178,4 , 83,5 , 871,3 , 76,7 , 416,7 , 708,0 , 46,8 , 39,0 , 349,3 , 566,5 , 80,6 , 188,8 , 41,3 , 174,8 , 394,8 , 4,2 , 146,6 , 564,0 , 5,3 , 242,5 .
10. X – godišnja količina oborina (litara/m^2): 634, 655, 483, 733, 679, 719, 471, 691, 621, 618, 567, 505, 541, 578, 540, 636, 571, 578, 525, 526, 557, 497, 589, 613, 652.
11. X – težina novorođenčeta (kg): 3,75 , 3,83 , 3,60 , 3,61 , 3,68 , 3,95 , 3,23 , 3,64 , 3,88 , 3,39 , 3,22 , 4,03 , 3,62 , 3,79 , 4,21 , 3,85 , 3,17 , 4,00 , 3,52 , 3,39 , 3,75 , 4,24 , 3,95 , 2,46 , 2,95 , 3,27 , 3,57 , 2,53 , 3,50 , 3,27 , 3,29 , 2,26 , 3,43 , 3,09 , 3,62 , 3,08 , 4,32 , 3,35 , 3,18 , 2,84 .
12. X – visina dvadesetogodišnjaka (cm): 185, 188, 177, 172, 180, 172, 175, 172, 179, 182, 169, 176, 160, 178, 176, 174, 170, 186, 172, 176, 169, 179, 182, 179, 165, 176, 159, 168, 174, 189, 182, 183, 181, 170, 168, 160, 178, 171, 174, 187, 166, 172.
13. X – tlačna čvrstoća cementne kocke (MPa): 27,8 , 29,5 , 22,8 , 23,5 , 24,7 , 23,1 , 26,6 , 24,5 , 26,7 , 24,9 , 28,4 , 29,6 , 22,0 , 26,1 , 24,8 , 24,5 , 27,6 , 23,1 , 24,6 , 23,5 , 23,2 , 26,0 , 25,4 , 23,3 , 23,7 , 26,7 , 23,1 , 24,5 , 24,5 , 23,4 , 24,6 , 22,2 , 21,5 , 26,5 , 23,5 , 23,3 , 26,6 , 27,6 , 26,9 , 22,1 .
14. X – vrijeme utrošeno za popravak stroja (sati): 6,07 , 1,09 , 3,67 , 0,35 , 0,68 , 0,06 , 0,51 , 0,55 , 0,46 , 4,24 , 0,80 , 2,21 , 0,77 , 0,96 , 6,28 , 3,67 , 1,72 , 0,64 , 2,40 , 1,60 , 0,19 , 2,12 , 1,30 , 6,14 , 3,18 .
15. Promatrajte početna slova svih riječi na str. 17 ove knjige, načinite pripadnu tablicu frekvencija i skicirajte odgovarajući histogram frekvencija. Usporedite dobivenu sliku sa sl. 6. Nacrtajte i sliku koja će biti analogna sl. 7.

II. Parametri niza statističkih podataka

1. Aritmetička sredina

Kada je riječ o iole brojnijem nizu statističkih podataka, onda već tablični, a pogotovo grafički prikaz tih podataka omogućuje vrlo jasan i pregledan uvid u bitna svojstva pojave koja se proučava uz pomoć tih podataka. No, odmah se može postaviti i zahtjev da se bitna svojstva promatranoga statističkog obilježja X izraze još sažetije, tj. da se karakteriziraju uz pomoć jednog *parametra* (broja) ili više njih, koji će se, dakako, na određeni način definirati pomoću danog niza $x_1, \ldots, x_n$ statističkih podataka o obilježju X.

Jedan od najvažnijih parametara koji grubo pokazuje smještaj (lokaciju) danih statističkih podataka na brojevnoj osi jest ***aritmetička sredina*** ili ***prosjek*** danog niza brojčanih podataka. Obično se označava sa $\overline{x}$ i definira formulom

(1) $$\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \,.$$

Tako se, na primjer, za statističke podatke iz 1. primjera I. poglavlja dobiva

$$\overline{x} = \frac{1}{30}(1 + 4 + 2 + 3 + \ldots + 3 + 3 + 2) = \frac{81}{30} = 2,7 \,,$$

pa se može reći da je u danom nizu ocjena iz matematike aritmetička sredina ili prosjek 2,7. Kaže se još da je srednja ocjena iz matematike učenika dotičnog razreda 2,7.

Odmah se može primijetiti da parametar $\overline{x}$ ima apstraktni karakter, jer je jasno da se vrijednost 2,7 uopće ne može realizirati pri opažanju statističkog obilježja X (ocjena iz matematike).

Za statističke podatke iz 5. primjera I. poglavlja dobiva se

$$\overline{x} = \frac{1}{100}(30{,}97 + 42{,}63 + \ldots + 33{,}09 + 29{,}60) = 35{,}52 \,.$$

To pokazuje da je prosjek izmjerenih tlačnih čvrstoća betonskih kocki 35,52 MPa.

Pogleda li se položaj broja 2,7 na sl. 1. i sl. 2. u prvom poglavlju, odnosno položaj broja 35,52 na apscisnoj osi na sl. 8, 9, 10. i 11, odmah se vidi da prosjek danog niza statističkih podataka pokazuje lokaciju tih podataka na apscisnoj osi i može se uzeti kao određeni reprezentant svih brojeva danog niza statističkih podataka.

Da bi se što bolje shvatio smisao definicijske formule (1) te praktično značenje prosjeka $\overline{x}$ danog niza podataka, korisno je pogledati glavna teorijska svojstva tako definirane veličine.

Temeljno svojstvo prosjeka definiranog u (1) izraženo je relacijom

(2) $$(x_1 - \overline{x}) + \ldots + (x_n - \overline{x}) = \sum_{i=1}^{n} (x_i - \overline{x}) = 0 .$$

Valjanost jednadžbe (2) je očigledna, a njome se izražava da je zbroj svih odstupanja podataka od njihova prosjeka jednak nuli. To pokazuje u kojem smislu treba shvatiti $\overline{x}$ kao određenu sredinu danog niza statističkih podataka.

Druga važna osobina prosjeka $\overline{x}$ može se uočiti ako se promotri zbroj kvadrata odstupanja danih podataka od nekoga realnog broja c ($c \in \mathbf{R}$). Dobiva se

$$\sum_{i=1}^{n} (x_i - c)^2 = \sum_{i=1}^{n} (x_i - \overline{x} + \overline{x} - c)^2 =$$
$$= \sum_{i=1}^{n} (x_i - \overline{x})^2 + 2(\overline{x} - c) \sum_{i=1}^{n} (x_i - \overline{x}) + n(\overline{x} - c)^2 .$$

Budući da je $\sum\limits_{i=1}^{n} (x_i - \overline{x}) = 0$, konačno se dobiva

(3) $$\sum_{i=1}^{n} (x_i - c)^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 + n(\overline{x} - c)^2 \geq 0 .$$

Iz (3) se razabire da je

(4) $$\sum_{i=1}^{n} (x_i - c)^2 \geq \sum_{i=1}^{n} (x_i - \overline{x})^2$$

i da znak jednakosti u (4) vrijedi onda i samo onda ako je $c = \overline{x}$. To znači da je zbroj kvadrata odstupanja danih podataka od prosjeka $\overline{x}$ manji od zbroja kvadrata odstupanja danih podataka od bilo kojega drugog broja $c \neq \overline{x}$.

Prema tome, $\overline{x}$ kao određena sredina danih podataka ima svojstvo da je zbroj odstupanja od $\overline{x}$ jednak nuli, a zbroj kvadrata odstupanja od $\overline{x}$, kao određena nenegativna veličina, najmanji je.

Ako su statistički podaci o nekome diskretnom obilježju već sređeni u tablicu frekvencija (tabl. 1 u I. 1), onda se prosjek $\overline{x}$ može izračunati pomoću formule

(5) $$\overline{x} = \frac{1}{n}(a_1 f_1 + \ldots + a_r f_r) = \frac{1}{n} \sum_{j=1}^{r} a_j f_j ,$$

odnosno pomoću formule

(6) $$\overline{x} = a_1 p_1 + \ldots + a_r p_r = \sum_{j=1}^{r} a_j p_j .$$

Valjanost formule (5) proizlazi iz samog pojma frekvencije f_j podatka a_j u danom nizu $x_1, \ldots, x_n$, dok (6) slijedi iz (5) kada se uzme u obzir da je $p_j = \frac{1}{n} f_j$ ($j = 1, \ldots, r$). Tako se, na primjer, prosjek statističkih podataka iz 1. primjera u I. 1. može izračunati i iz tabl. 2 u I. 1. Primjenom (5) dobiva se

$$\overline{x} = \frac{1}{30}(1 \cdot 3 + 2 \cdot 11 + 3 \cdot 10 + 4 \cdot 4 + 5 \cdot 2) = 2{,}7,$$

a primjenom (6) proizlazi

$$\overline{x} = 1 \cdot \frac{1}{3} + 2 \cdot \frac{11}{30} + 3 \cdot \frac{1}{3} + 4 \cdot \frac{2}{15} + 5 \cdot \frac{1}{15} = 2{,}7.$$

Ako je riječ o podacima kontinuiranoga statističkog obilježja i ako se za izračunavanje prosjeka $\overline{x}$ želi iskoristiti tablica frekvencija gdje su podaci grupirani u razrede (tabl. 6. u I. 5), onda se može definirati veličina

(7) $$\tilde{x} = \frac{1}{n} \sum_{j=1}^{r} \overline{a}_j f_j = \sum_{j=1}^{r} \overline{a}_j p_j .$$

Veličina $\overline{x}$ definirana u (1) i veličina $\tilde{x}$ definirana u (7) općenito nisu jednake, jer je $\tilde{x}$ *aritmetička sredina* ili *prosjek grupiranih podataka*, pri čemu je izvorni podatak zamijenjen sredinom pripadnog razreda. Očigledno je da i parametar $\tilde{x}$ pokazuje položaj danih podataka na brojevnoj osi i kao takav se ubraja u tzv. *parametre lokacije*. Osim toga $\overline{x}$ i $\tilde{x}$ redovito se malo razlikuju, pa se za praktične potrebe često umjesto $\overline{x}$ uzima $\tilde{x}$.

Ako se, na primjer, na temelju tabl. 5. iz I. 4. izračuna

$$\tilde{x} = \frac{1}{100}(25{,}6 \cdot 2 + 26{,}8 \cdot 3 + \ldots + 47{,}2 \cdot 2 + 48{,}4 \cdot 0) = 35{,}49,$$

vidi se da postoji mala razlika od vrijednosti $\overline{x} = 35{,}52$ dobivene na temelju formule (1) i izvornog niza (negrupiranih) podataka.

Uzme li se, pak tabl. 7. iz I. 6. kao osnova za računanje prosjeka grupiranih podataka, dobiva se

$$\tilde{x}_1 = \frac{1}{100}(26{,}2 \cdot 5 + 28{,}6 \cdot 6 + \ldots + 47{,}8 \cdot 2) = 35{,}55,$$

pa se vidi da je dobivena vrijednost različita i od $\overline{x}$ i od $\tilde{x}$, ali da su te razlike praktički beznačajne.

Ako se razdioba relativnih frekvencija interpretira kao linijska razdioba ukupne mase 1, kako je to već ranije opisano u I. 8, onda se statistički pojam prosjeka može

mehanički interpretirati kao *težište*. Formule (6) i (7), kojima su definirane veličine $\bar{x}$ i $\tilde{x}$, identične su formulama kojima se u mehanici definira apscisa težišta za danu linijsku razdiobu mase.

Ako se još pretpostavi da se raspodijeljena masa nalazi u određenome gravitacijskom polju, onda su uz mase vezane i odgovarajuće sile, pa se veličina $\bar{x}$, odnosno $\tilde{x}$, može interpretirati i kao ***statički moment*** tih sila u odnosu na ishodišnu točku brojevne osi. Promotri li se statički moment tih sila u odnosu na točku apscise $\bar{x}$ (težište), dobiva se

$$(a_1 - \bar{x})p_1 + \ldots + (a_r - \bar{x})p_r = 0,$$

a to je poznati rezultat iz mehanike da je statički moment danih sila u odnosu na težište jednak nuli.

Može se, prema tome, reći da prosjek danog niza statističkih podataka reprezentira cijeli niz podataka na isti način kao što težište odgovarajuće linijske razdiobe masa reprezentira ukupnu masu.

2. Medijan

Osim prosjeka niza statističkih podataka definiraju se još i neki drugi parametri lokacije. Najpoznatiji je ***medijan***, koji ćemo označivati sa m. Za definiciju medijana m pretpostavit ćemo da su podaci poredani po veličini, tj. da vrijedi $x_1' \leq x_2' \leq \ldots \leq x_n'$. Tada se definira

$$m = \begin{cases} \frac{1}{2}(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}), & \text{za parno } n \\ x'_{\frac{n+1}{2}}, & \text{za neparno } n. \end{cases} \tag{8}$$

Zanimljivo je primijetiti da na vrijednost medijana m utječu samo središnji podaci iz danoga uređenog niza $x_1', \ldots, x_n'$, za razliku od prosjeka $\bar{x}$ na kojega svaki od podataka ima određeni utjecaj. Medijan se neće izmijeniti ako se, recimo, najmanji od danih n ($n \geq 3$) podataka proizvoljno smanji, ili najveći podatak po volji poveća. Medijan m je neosjetljiv na ekstremne vrijednosti u danom nizu podataka.

U 1. primjeru iz I. 1. imali smo $n = 30$ i niz danih podataka poredan po veličini izgleda ovako:

$$x_1' = 1, x_2' = 1, x_3' = 1, x_4' = 2, x_5' = 2, x_6' = 2, x_7' = 2, x_8' = 2, x_9' = 2, x_{10}' = 2,$$
$$x_{11}' = 2, x_{12}' = 2, x_{13}' = 2, x_{14}' = 2, x_{15}' = 3, x_{16}' = 3, x_{17}' = 3, x_{18}' = 3, x_{19}' = 3, \ldots$$

Stoga iz (8) proizlazi da je

$$m = \frac{1}{2}(x_{15}' + x_{16}') = 3.$$

Za statističke podatke iz tabl. 3. u I. 2. vidljivo je da je $n = 200$, $x'_{\frac{n}{2}} = x'_{100} = 10$ i $x'_{\frac{n}{2}+1} = x'_{101} = 10$, pa se na temelju (8) dobiva

$$m = \frac{1}{2}(x'_{100} + x'_{101}) = 10.$$

Praktična interpretacija medijana m, kao određenog parametra lokacije danog niza statističkih podataka, sastoji se u tome da je m ona točka na brojevnoj osi, koja ima svojstvo da se lijevo i desno od nje nalazi jednak broj podataka danog niza. Točka m, dakle, dijeli dani niz podataka na dva jednakobrojna dijela.

Medijan ima još jedno zanimljivo svojstvo, a to je da zbroj apsolutnih vrijednosti odstupanja danih podataka od nekog broja c ($c \in \mathbf{R}$) poprima minimalnu vrijednost, ako se uzme $c = m$. Dokažimo to. Očigledno se može pisati

$$\sum_{i=1}^{n} |x_i' - c| = \begin{cases} \sum_{i=1}^{\frac{n}{2}} (|x_i' - c| + |x'_{n+1-i} - c|), & \text{za parno } n \\ \sum_{i=1}^{\frac{n-1}{2}} (|x_i' - c| + |x'_{n+1-i} - c|) + |x'_{\frac{n+1}{2}} - c|, & \text{za neparno } n. \end{cases}$$

Uzme li se $c = m$, vidi se da je

$$|x_i' - m| + |x'_{n+1-i} - m| = |x_i' - x'_{n+1-i}|,$$

za $i = 1, \ldots, \frac{n}{2}$ kada je n parno, odnosno za $i = 1, \ldots, \frac{n-1}{2}$ kada je n neparno. Iz ovoga odmah slijedi da je

$$\sum_{i=1}^{n} |x_i' - m| = \begin{cases} \sum_{i=1}^{\frac{n}{2}} |x_i' - x'_{n+1-i}|, & \text{za parno } n \\ \sum_{i=1}^{\frac{n-1}{2}} |x_i' - x'_{n+1-i}|, & \text{za neparno } n. \end{cases}$$

Ako je n parno i $c \neq m$, te ako između c i m nema nijednog podatka danog niza, onda je

$$\sum_{i=1}^{n} |x_i' - c| = \sum_{i=1}^{n} |x_i' - m|.$$

Ako je $c \neq m$, recimo $c < m$, i bar jedan od podataka nalazi se u intervalu $[c, m]$, tada za $x_i' \leq c$ vrijedi

$$\begin{aligned} |x_i' - c| + |x'_{n-i+1} - c| &= |x_i' - x'_{n-i+1}| + 2|x_i' - c| = \\ &= |x_i' - m| + |x'_{n-i+1} - m| + 2|x_i' - c|. \end{aligned}$$

Za $c < x_i < m$ vrijedi

$$|x_i' - c| + |x'_{n-i+1} - c| = |x_i' - x'_{n-i+1}| = |x_i' - m| + |x'_{n-i+1} - m|.$$

Stoga za $c < m$ vrijedi

$$\sum_{i=1}^{n} |x_i' - c| = \sum_{i=1}^{\frac{n}{2}} |x_i' - x'_{n-i+1}| + 2\sum_{i=1}^{\frac{n}{2}} |x_i' - c| \geq \sum_{i=1}^{n} |x_i' - m|.$$

Slično se analizira i slučaj $c > m$, a također i slučaj neparnoga n. U svakom se slučaju dobiva

$$(9) \qquad \sum_{i=1}^{n} |x'_i - c| \geq \sum_{i=1}^{n} |x'_i - m|,$$

pri čemu znak jednakosti vrijedi kada je $c \in [x'_{\frac{n}{2}}, x'_{\frac{n}{2}+1}]$, za parno n, odnosno $c = m = x'_{\frac{n+1}{2}}$, za neparno n.

Budući da konačna suma ne ovisi o redoslijedu svojih članova, (9) se može pisati i kao

$$(10) \qquad \sum_{i=1}^{n} |x_i - c| \geq \sum_{i=1}^{n} |x_i - m|,$$

čime je dokazana izrečena tvrdnja o svojstvu medijana.

3. Varijanca

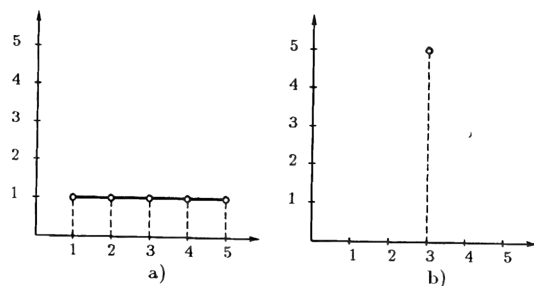
Parametrima lokacije očigledno nije obuhvaćen jedan vrlo značajan i važan aspekt danog niza $x_1, \ldots, x_n$ statističkih podataka, a to je njihovo rasipanje ili raspršenje. Da bi se lakše shvatio smisao i potreba definiranja parametra koji će pokazivati raspršenje danog niza podataka, dobro je uočiti dva vrlo tipična slučaja. Zamislimo da su mjerenjem veličine X dobiveni ovi rezultati:

$$(x) \qquad x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4, x_5 = 5,$$

a mjerenjem veličine Y ovi rezultati:

$$(y) \qquad y_1 = y_2 = y_3 = y_4 = y_5 = 3.$$

Oba niza (x) i (y) imaju jednake prosjeke ($\overline{x} = \overline{y} = 3$), a također i jednake medijane ($m_x = m_y = 3$). Odmah se, međutim, vidi da u nizu (y) uopće nema raspršenja podataka, već su svi podaci koncentrirani u jednoj točki, dok je u nizu (x) znatno raspršenje podataka.



Slika 15. Poligoni frekvencija za nizove (x) i (y)

Stoga se prirodno nameće ideja da se definira jedan parametar ili više njih, koji će "mjeriti" rasipanje podataka. Iako će se kasnije definirati i neki drugi parametri koji pokazuju rasipanje podataka, najprije će se definirati i razmotriti svojstva najvažnijeg parametra rasipanja koji se zove ***varijanca*** ili ***disperzija*** danog niza statističkih podataka.

Varijanca se označuje sa s_0^2 i definira formulom

$$(11) \qquad s_0^2 = \frac{1}{n}[(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Iz formule (11) odmah se vidi da je s_0^2 zbroj kvadrata odstupanja pojedinih podataka od prosjeka $\overline{x}$ podijeljen s brojem n svih podataka, pa se može reći i da je s_0^2 aritmetička sredina kvadrata odstupanja od prosjeka u danom nizu statističkih podataka. Također se govori da je s_0^2 prosječno kvadratno odstupanje od prosjeka.

Nenegativna veličina $s_0 = \sqrt{s_0^2}$ zove se ***standardno odstupanje*** ili ***standardna devijacija*** danog niza podataka.

Očigledno je $s_0^2 \geq 0$, pri čemu može biti $s_0^2 = 0$ onda i samo onda ako je $x_1 = \ldots = x_n = \overline{x}$, tj. kada svi podaci padaju u istu točku. Takvu situaciju imamo u nizu (y), što se zorno vidi na sl. 15b, pa se može reći da je varijanca podataka niza (y) jednaka nuli. Za podake niza (x) (sl. 15a) dobiva se

$$s_0^2 = \frac{1}{5}[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2] = 2,$$

odnosno

$$s_0 = \sqrt{2} \approx 1{,}41.$$

Ako su podaci o diskretnome statističkom obilježju već sređeni u obliku tablice frekvencija (tabl. 1. u I. 1), onda se definicijska formula za varijancu može zapisati i kao

$$(12) \qquad s_0^2 = \frac{1}{n}\sum_{j=1}^{r}(a_j - \overline{x})^2 f_j = \sum_{j=1}^{r}(a_j - \overline{x})^2 p_j.$$

Valjanost formule (12) izlazi iz samog pojma frekvencije f_j, odnosno relativne frekvencije p_j, vrijednosti $a_j \in A$ (A je skup mogućih vrijednosti promatranoga diskretnog statističkog obilježja).

Za podatke iz 1. primjera u I. 1, sređene u tabl. 2. iz I. 1, primjenom formule (12) dobiva se

$$\begin{aligned} s_0^2 &= \frac{1}{30}[(1-2{,}7)^2 \cdot 3 + (2-2{,}7)^2 \cdot 11 + (3-2{,}7)^2 \cdot 10 + \\ &+ (4-2{,}7)^2 \cdot 4 + (5-2{,}7)^2 \cdot 2] = \\ &= (1-2{,}7)^2 \cdot \frac{1}{10} + (2-2{,}7)^2 \cdot \frac{11}{30} + (3-2{,}7)^2 \cdot \frac{1}{3} + \\ &+ (4-2{,}7)^2 \cdot \frac{1}{15} + (5-2{,}7)^2 \cdot \frac{1}{15} \approx 1{,}08, \end{aligned}$$

odnosno

$$s_0 = \sqrt{1{,}08} \approx 1{,}04.$$

Za podatke iz 3. primjera u I. 2. imamo $n = 200$, $r = 20$ i $\overline{x} = 9{,}97$, tako da se primjenom (12) dobiva

$$s_0^2 = \frac{1}{200}[(3-9{,}97)^2 \cdot 1 + (4-9{,}97)^2 \cdot 2 + \ldots + (22-9{,}97)^2 \cdot 2] \approx 9{,}64,$$
$$s_0 = \sqrt{9{,}64} \approx 3{,}10.$$

Ako je riječ o podacima sređenim u tablicu frekvencija, pri čemu su podaci grupirani u razrede (tabl. 6. u I. 5), onda se definira veličina

$$\bar{s}_0^2 = \frac{1}{n}\sum_{j=1}^{r}(\overline{a}_j - \overline{x})^2 f_j = \sum_{j=1}^{r}(\overline{a}_j - \overline{x})^2 p_j, \tag{13}$$

koja se zove *varijanca grupiranih podataka*. Veličina $\bar{s}_0^2$ općenito neće biti jednaka varijanci s_0^2 definiranoj formulom (11), jer ona zapravo pokazuje rasipanje sredina razreda, "opterećenih" pripadnim frekvencijama, oko prosjeka $\overline{x}$. Veličine s_0^2 i $\bar{s}_0^2$ redovito se malo razlikuju, tako da se u praksi katkada s_0^2 zamjenjuje sa $\bar{s}_0^2$.

Ako se u formuli (11) izvrši naznačeno kvadriranje, dobiva se

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\overline{x}\sum_{i=1}^{n}x_i + n\overline{x}\right).$$

Budući da je $\sum_{i=1}^{n} x_i = n\overline{x}$, konačno se dobiva formula

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2. \tag{14}$$

Za numeričko računanje varijance često je formula (14) prikladnija od formule (11). Formula (14) kazuje da je varijanca jednaka razlici aritmetičke sredine kvadrata i kvadrata aritmetičke sredine (prosjeka) danih podataka.

Definiciju varijance kao određene mjere rasipanja danog niza statističkih podataka opravdat će i svojstvo varijance koje ćemo sada razmotriti. Najprije valja uočiti da nije moguće dobiti mjeru rasipanja ako se promatra zbroj odstupanja (bez kvadrata) od prosjeka. Prosjek $\overline{x}$ je, naime, upravo tako definiran da se pozitivna i negativna odstupanja međusobno ponište, što je izraženo relacijom (2). Kvadrati u formuli (11) imaju upravo tu ulogu da dođu do izražaja i pozitivna i negativna odstupanja te da veća odstupanja (veća od jedan) više utječu na vrijednost varijance.

Uzme li se proizvoljan realan broj c, tada iz (3) proizlazi

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 + (\overline{x} - c)^2 = s_0^2 + (\overline{x} - c)^2 \geq s_0^2. \tag{15}$$

Iz (15) se vidi da vrijednost izraza $\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$, tj. aritmetička sredina kvadrata odstupanja danih podataka od broja c, ovisi o broju c tako da poprima minimalnu vrijednost za $c = \overline{x}$ i tada je njegova vrijednost upravo varijanca s_0^2. Pojednostavnjeno se može reći da je rasipanje podataka, mjereno kvadratnim odstupanjima, minimalno ako se uzimaju odstupanja od prosjeka tih podataka. Kvadratna odstupanja podataka od bilo kojeg drugog broja $c \neq \overline{x}$ veća su od varijance s_0^2.

4. Standardna i apsolutna devijacija

Značenje standardne devijacije s_0, kao određenog parametra rasipanja, može se vidjeti iz sljedećeg rezoniranja. Iz formule (11) proizlazi da je

$$n\,s_0^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2 = \sum_{x_i<\overline{x}-ks_0}(x_i-\overline{x})^2 + \sum_{\overline{x}-ks_0\leq x_i\leq\overline{x}+ks_0}(x_i-\overline{x})^2 + \sum_{x_i>\overline{x}+ks_0}(x_i-\overline{x})^2.$$

Tu je zbroj po i-ovima od 1 do n razbijen na tri nenegativna pribrojnika. U prvom se zbraja po onim i-ovima za koje je $x_i < \overline{x} - ks_0$, gdje je $k > 0$ proizvoljan broj, u drugom po onim i-ovima za koje je $\overline{x} - ks_0 \leq x_i \leq \overline{x} + ks_0$ i u trećem po onima za koje je $x_i > \overline{x} + ks_0$. Stoga se može pisati

$$ns_0^2 \geq \sum_{x_i<\overline{x}-ks_0}(x_i-\overline{x})^2 + \sum_{x_i>\overline{x}+ks_0}(x_i-\overline{x})^2 = \sum_{|x_i-\overline{x}|>ks_0}(x_i-\overline{x})^2.$$

Ako se sada u prvi zbroj umjesto x_i stavi $\overline{x} - ks_0$, a u drugi se umjesto x_i stavi $\overline{x} + ks_0$, njihova će se vrijednost smanjiti, tj. vrijedi

$$\sum_{|x_i-\overline{x}|\geq ks_0}(x_i-\overline{x})^2 \geq \sum_{x_i<\overline{x}-ks_0}k^2s_0^2 + \sum_{x_i>\overline{x}+ks_0}k^2s_0^2 = \sum_{|x_i-\overline{x}|>ks_0}k^2s_0^2.$$

Ako u danom nizu podataka $x_1, \ldots, x_n$ ima njih $l(l \leq n)$ za koje vrijedi $|x_i - \overline{x}| > ks_0$, onda se može pisati

$$ns_0^2 \geq lk^2s_0^2,$$

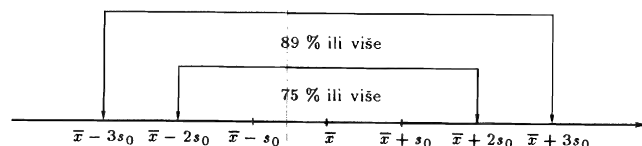
pa ako je još $s_0 > 0$, dobiva se

$$l \leq \frac{1}{k^2}n. \tag{16}$$

Specijalno za $k = 3$, dobiva se da je $l \leq \frac{1}{9}n$, a to znači da se najviše $\frac{1}{9}$ ili oko 11 % svih n podataka nalazi izvan intervala $[\overline{x} - 3s_0, \overline{x} + 3s_0]$, odnosno da se bar 89 % svih podataka nalazi unutar intervala širine $6s_0$ sa središtem u $\overline{x}$.

Prema tome, poznavanje samo dvaju parametara $\overline{x}$ i s_0 danog niza statističkih podataka omogućuje nam već vrlo dobar uvid u položaj i rasipanje danih podataka. Tako, na primjer, za $k = 2$ relacija (16) pokazuje da je $l \leq \frac{1}{4}n$, što znači da se

unutar intervala $[\overline{x} - 2s_0, \overline{x} + 2s_0]$ nalazi bar 75 % ukupnog broja podataka danog niza $x_1, \ldots, x_n$.



Slika 16. Skica razdiobe podataka na brojevnoj osi

Rasipanje podataka može se kvantitavno opisati i pomoću apsolutnih vrijednosti odstupanja od medijana m. Iz (10), naime, proizlazi da je

$$(17)\qquad \frac{1}{n}\sum_{i=1}^{n}|x_i - c| \geq \frac{1}{n}\sum_{i=1}^{n}|x_i - m|,$$

što pokazuje da aritmetička sredina apsolutnih vrijednosti odstupanja podataka od medijana m nije veća od aritmetičke sredine apsolutnih vrijednosti odstupanja podataka od bilo kojega drugog realnog broja c. Stoga se veličina

$$(18)\qquad a = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|$$

također može uzeti kao određena mjera rasipanja danog niza statističkih podataka. Veličina a zove se *apsolutna devijacija oko medijana*.

Ako se radi s frekvencijama, odnosno relativnim frekvencijama, onda formula (18) postaje

$$(19)\qquad a = \frac{1}{n}\sum_{j=1}^{r}|a_j - m|f_j = \sum_{j=1}^{r}|a_j - m|p_j.$$

Očigledno je $a \geq 0$ i znak jednakosti vrijedi onda i samo onda ako vrijedi $x_1 = \ldots = x_n = m$, tj. i ova mjera rasipanja poprima vrijednost nula za podatke koji padaju u istu točku.

U II. 2. izračunan je medijan $m = 1{,}5$ za podatke iz 1. primjera u I. 1, pa se vidi da je za te podatke apsolutna devijacija oko medijana

$$a = \frac{1}{30}[|1-3|3 + |2-3|11 + |3-3|10 + |4-3|4 + |5-3|2] = \frac{5}{6} \approx 0{,}83.$$

Usporedi li se dobivena vrijednost apsolutne devijacije oko medijana sa standardnom devijacijom istih podataka izračunanom u prethodnom poglavlju ($s_0 \approx 1{,}04$), vidi se da je u ovom primjeru $a < s_0$. Istaknuli smo već i slučaj kada je $a = s_0$, a može biti i $a > s_0$.

5. Raspon i interkvartilni raspon

Ako su podaci poredani po veličini, tj. $x'_1 \leq \ldots \leq x'_n$ $(n \geq 2)$ onda se kao određeni pokazatelj rasipanja može uzeti razlika d između najveće x'_n i najmanje x'_1 vrijednosti u danom nizu statističkih podataka. Piše se

$$(20)\qquad d = \max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\} = x'_n - x'_1$$

i veličina d zove se ***raspon*** niza $x_1, \ldots, x_n$ statističkih podataka.

Očigledno je $d \geq 0$ i znak jednakosti vrijedi onda i samo onda ako je $x_1 = \ldots = x_n$.

Za podatke iz 1. primjera u I. 1. vidi se da je $x'_1 = 1$ i $x'_{30} = 5$, tako da je $d = 4$.

Pogledaju li se podaci u 5. primjeru iz I. 4. vidi se da je $x'_1 = x_{84} = 25{,}13$ i $x'_{100} = x_7 = 47{,}12$, tako da je odgovarajući raspon $d = x'_{100} - x'_1 = 21{,}98$.

Glavni nedostatak raspona kao određene mjere rasipanja podataka sastoji se u tome što raspon ovisi samo o ekstremnim vrijednostima u danom nizu statističkih podataka, dok središnja skupina podataka ne utječe na veličinu raspona. To znači da vrlo različiti nizovi podataka mogu imati istu vrijednost raspona.

Tako, na primjer, već navedeni raspon $d = 21{,}98$ ima niz od $n = 100$ podataka iz 5. primjera u I. 4, ali isto tako i niz od samo dva podatka $y_1 = x'_1 = 25{,}13$ i $y_2 = x'_{100} = 47{,}12$. Usporede li se, recimo, njihove standardne devijacije vidi se da je standardna devijacija podataka iz 5. primjera u I. 4. $s_0 = 4{,}63$, dok za niz od samo dva ekstremna podatka y_1 i y_2 standardna devijacija iznosi $s_0 = 11$.

Da bi se donekle uklonili navedeni nedostaci raspona kao pokazatelja rasipanja a zadržala njegova dobra svojstva, definira se tzv. interkvartilni raspon. Grubo rečeno, interkvartilni raspon je raspon niza koji se dobije od zadanog niza podataka kada se odbaci četvrtina najmanjih i četvrtina najvećih vrijednosti iz danog niza statističkih podataka. Ako je, dakle, n $(n \geq 4)$ djeljiv sa 4 i vrijedi

$$x'_1 \leq \ldots \leq x'_{\frac{n}{4}} \leq x'_{\frac{n}{4}+1} \ldots \leq x'_{\frac{3n}{4}} \leq x'_{\frac{3n}{4}+1} \leq \ldots \leq x'_n,$$

onda se veličina

$$(21)\qquad d_2 = x'_{\frac{3n}{4}} - x'_{\frac{n}{4}+1}$$

zove ***interkvartilni raspon*** danog niza statističkih podataka.

Za podatke iz 1. primjera u I. 1. interkvartilni raspon odredio bi se tako da se, zbog $n = 30$, što nije djeljivo sa 4, iz niza izbace bilo koja dva podatka, recimo $x_{29} = 0$ i $x_{30} = 2$, pa se preostali niz uredi po veličini. Dobiva se niz u kojem $n = 28$, što je djeljivo sa 4 i $x'_{\frac{n}{4}+1} = x'_8 = 2$, $x'_{\frac{3n}{4}} = x'_{21} = 3$, tako da je pripadni interkvartilni raspon $d_2 = 1$.

6. Parametri oblika

Sada se, prirodno, nameće ideja da se definiraju još neki parametri koji će karakterizirati i druga svojstva niza statističkih podataka $x_1, \ldots, x_n$. Općenito se definiraju tzv. ***statistički momenti***, tako da se stavi

(22) $$b_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \qquad k = 0, 1, \ldots$$

Parametar b_k zove se *ishodišni* ili *pomoćni moment k-tog reda*, a parametar

(23) $$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k, \qquad k = 0, 1, \ldots$$

zove se ***centralni*** ili ***glavni moment k-tog reda*** danog niza statističkih podataka.

Očigledno je $b_0 = m_0 = 1$, $b_1 = \overline{x}$, $m_1 = 0$ i $m_2 = s_0^2$.

Iz (22) i (23) proizlazi da je

(24) $$m_2 = b_2 - b_1^2,$$

(25) $$m_3 = b_3 - 3b_1b_2 + 2b_1^3,$$

(26) $$m_4 = b_4 - 4b_3b_1 + 6b_2b_1^2 - 3b_1^4,$$

pa se vidi da formule (24), (25) i (26) omogućuju da se centralni momenti m_2, m_3 i m_4 izraze pomoću ishodišnih momenata, koji su definirani jednostavnijim formulama.

Uloga centralnog momenta trećeg reda m_3 može se nazreti iz ovog rezoniranja. Ako su podaci $x_1, \ldots, x_n$ raspoređeni simetrično oko točke $\overline{x}$, onda svakoj vrijednosti x_i odgovara simetrična vrijednost x_i', tako da je $x_i - \overline{x} = -(x_i' - \overline{x})$ i $(x_i - \overline{x})^3 = -(x_i' - \overline{x})^3$, što ima za posljedicu da je $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 = 0$. Ako je $m_3 > 0$, onda to znači da je $\sum_{i=1}^{n}(x_i - \overline{x})^3 > 0$, tj. podaci su "razvučeniji" desno od $\overline{x}$, odnosno "zbijeniji" su lijevo od $\overline{x}$, a ako je $m_3 < 0$, onda su podaci "razvučeniji" lijevo od $\overline{x}$, a "zbijeniji" desno od $\overline{x}$.

Uobičajeno je da se parametar K definiran formulom

(27) $$K = \frac{m_3}{s_0^3}$$

zove ***koeficijent asimetrije*** danog niza statističkih podataka. Za simetrično raspoređene podatke oko prosjeka $\overline{x}$ je $K = 0$. Ako je $K < 0$ govori se o negativnoj, a ako je $K > 0$ o pozitivnoj asimetriji danog niza statističkih podataka. Budući da se simetričnost ili asimetričnost podataka zorno očituje u obliku pripadnog poligona, odnosno histograma, frekvencija, parametar K pripada u tzv. *parametre oblika*.

Kao drugi parametar oblika definira se ***koeficijent spljoštenosti*** E, i to formulom

(28) $$E = \frac{m_4}{s_0^4} - 3.$$

Ako je $E = 0$, onda se govori o tzv. *normalnoj spljoštenosti* grafikona frekvencija danog niza podataka, a ako je $E \neq 0$, onda se E interpretira kao odstupanje (*eksces*) od normalnosti. Dakako, eksces može biti pozitivan i negativan.

Ako su podaci sređeni u tablici frekvencija, onda se za proračun statističkih momenata, umjesto formula (22) i (23), mogu primijeniti formule

(29) $$b_k = \frac{1}{n}\sum_{j=1}^{r} a_j^k f_j = \sum_{j=1}^{r} a_j^k p_j,$$

(30) $$m_k = \frac{1}{n}\sum_{j=1}^{r}(a_j - \overline{x})^k f_j = \sum_{j=1}^{r}(a_j - \overline{x})^k p_j.$$

Za podatke iz 1. primjera u I. 1. proračun koeficijenta asimetrije K i koeficijenta spljoštenosti E može se provesti na sljedeći način:

a_j	f_j	$a_j f_j$	$a_j^2 f_j$	$a_j^3 f_j$	$a_j^4 f_j$
1	3	3	3	3	3
2	11	22	44	88	176
3	10	30	90	270	810
4	4	16	64	256	1 024
5	2	10	50	250	1 250
$\sum$	30	81	251	867	3 263

$$b_1 = \frac{1}{30}\cdot 81 = 2{,}7 = \overline{x},$$

$$b_2 = \frac{1}{30}\cdot 251 \approx 8{,}37, \quad s_0^2 = m_2 = b_2 - b_1^2 \approx 1{,}08, \quad s_0 \approx 1{,}04,$$

$$b_3 = \frac{1}{30}\cdot 867 \approx 28{,}9, \quad m_3 = b_3 - 3b_1b_2 + 2b_1^3 = 0{,}47, \quad K = \frac{m_3}{s_0^3} \approx 0{,}42,$$

$$b_4 = \frac{1}{30}\cdot 3\,263 \approx 108{,}77, \quad m_4 = b_4 - 4b_3b_1 + 6b_2b_1^2 - 3b_1^4 \approx 3{,}32,$$

$$E = \frac{m_4}{s_0^4} - 3 = -0{,}15.$$

Primjedba

Primjena osobnog računala i odgovarajućeg programskog paketa omogućuje da se za dani niz statističkih podataka "pritiskom na gumb" dobiju vrijednosti najvažnijih parametara, kao što su prosjek, medijan, varijanca, standardna devijacija, apsolutna devijacija, minimalna vrijednost, maksimalna vrijednost, raspon, interkvartilni raspon, koeficijent asimetrije i koeficijent spljoštenosti.

Zadaci

1. Za svaki niz statističkih podataka naveden u zadacima 1–14. u I. poglavlju izračunajte:

a) prosjek,
b) medijan,
c) varijancu,
d) standardnu devijaciju,
e) apsolutnu devijaciju,
f) raspon,
g) interkvartilni raspon,
h) koeficijent asimetrije,
i) koeficijent spljoštenosti.

2. Neka je $x_1, \dots, x_n$ niz statističkih podataka s prosjekom $\overline{x}$ i varijancom s_x^2, $A \neq 0$ i B zadani realni brojevi, te $y_i = Ax_i + B$, $i = 1, \dots, n$. Dokažite da nizu $y_1, \dots, y_n$ pripada prosjek $\overline{y} = A\overline{x} + B$ i varijanca $s_y^2 = A^2 s_x^2$.

3. Svaki niz statističkih podataka $x_1, \dots, x_n$ može se transformacijom oblika $y_i = x_i + B$ pretvoriti u niz $y_1, \dots, y_n$ prosjeka nula. Dokažite!

4. Na temelju n_1 mjerenja $x_1', \dots, x_{n_1}'$ statističkog obilježja X dobiven je prosjek $\overline{x}_1$ i varijanca s_1^2, a na temelju novih n_2 mjerenja $x_1'', \dots, x_{n_2}''$ dobiven je prosjek $\overline{x}_2$ i varijanca s_2^2. Ako se svih $n_1 + n_2 = n$ mjerenja shvate kao jedan niz statističkih podataka, onda se dobiva prosjek $\overline{x}$ i varijanca s_0^2. Dokažite da vrijedi:

a) $\overline{x} = \frac{1}{n}(n_1\overline{x}_1 + n_2\overline{x}_2)$,

b) $s_0^2 = \frac{1}{n}\{n_1[s_1^2 + (\overline{x}_1 - \overline{x})^2] + n_2[s_2^2 + (\overline{x}_2 - \overline{x})^2]\}$.

5. Dokažite da za podatke o diskretnom obilježju sa skupom vrijednosti $A = \{a_1, \dots, a_r\}$ vrijedi formula

$$s_0^2 = \frac{1}{n}\sum_{j=1}^{r} a_j^2 f_j - \overline{x}^2.$$

6. Dokažite da je standardna devijacija niza od dva različita podatka jednaka polovini pripadnog raspona.

7. Izvedite formule (24), (25) i (26).

III. Statistički podaci o dvodimenzionalnom obilježju

1. Kontingencijska tablica

Da bi se neka pojava proučila, često nije dovoljno promatrati samo jednu veličinu, već je nužno simultano promatrati više veličina i ustanoviti eventualnu ovisnost među tim veličinama. Budući da se neke opće ideje za simultano proučavanje više statističkih obilježja mogu uočiti već pri promatranju dvaju obilježja, detaljnije ćemo razmotriti probleme u vezi s tretiranjem statističkih podataka o dva statistička obilježja X i Y koja se simultano promatraju. To znači da se višestrukim ponavljanjem mjerenja, odnosno opažanja dotične pojave, dobiva niz uređenih parova realnih brojeva $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Govori se još da se promatra *dvodimenzionalno statističko obilježje* (X, Y) i da je dobiveni niz uređenih parova realnih brojeva odgovarajući niz statističkih podataka za dvodimenzionalno statističko obilježje (X, Y).

1. primjer

U jednom razredu od $n = 30$ učenika promatra se ocjena iz matematike (X) i ocjena iz fizike (Y), kao dvodimenzionalno statističko obilježje (X, Y). Uvidom u "imenik" dobiveni su ovi rezultati:

(1,3), (4,3), (2,2), (3,2), (1,2), (1,1), (2,2), (4,4), (2,2), (3,3), (4,4), (5,5), (3,5), (2,1), (2,3), (2,2), (5,5), (3,3), (2,2), (2,2), (3,3), (3,2), (4,4), (2,2), (3,3), (2,1), (3,2), (3,2), (3,2), (2,2).

Vidi se da je prvi učenik imao ocjenu $x_1 = 1$ iz matematike i ocjenu $y_1 = 3$ iz fizike, drugi učenik je imao ocjenu $x_2 = 4$ iz matematike i ocjenu $y_2 = 3$ iz fizike itd.

Iz danog niza statističkih podataka (uređenih parova) vidljivo je da se određeni uređeni parovi pojavljuju u tom nizu i više puta. Tako se, na primjer, uočava da se uređeni par (2,2) pojavljuje 8 puta, uređeni par (3,2) pojavljuje se 5 puta itd. Govori se još da uređenom paru (2,2) pripada frekvencija 8, odnosno relativna frekvencija $\frac{8}{30} \approx 0{,}27$.

Da bi se dobio jasniji i pregledniji uvid u dani niz statističkih podataka o promatranome dvodimenzionalnom obilježju (X, Y), prikladno je načiniti tablicu frekvencija (tabl. 1) i tablicu relativnih frekvencija (tabl. 2) za dane podatke.

Na gornjoj margini tabl. 1. i 2. upisane su vrijednosti obilježja Y, a na lijevoj margini vrijednosti obilježja X. U oba slučaja to su brojevi 1, 2, 3, 4 i 5, tj. moguće ocjene iz fizike, odnosno matematike. U nutarnja polja tabl. 1. upisane su odgo-

Tablica 1.

X \ Y	1	2	3	4	5	$\sum$
1	1	1	1	–	–	3
2	2	8	1	–	–	11
3	–	5	4	–	1	10
4	–	–	1	3	–	4
5	–	–	–	–	2	2
$\sum$	3	14	7	3	3	30

Tablica 2.

X \ Y	1	2	3	4	5	$\sum$
1	$\frac{1}{30}$	$\frac{1}{30}$	$\frac{1}{30}$	–	–	$\frac{3}{30}$
2	$\frac{2}{30}$	$\frac{8}{30}$	$\frac{1}{30}$	–	–	$\frac{11}{30}$
3	–	$\frac{5}{30}$	$\frac{4}{30}$	–	$\frac{1}{30}$	$\frac{10}{30}$
4	–	–	$\frac{1}{30}$	$\frac{3}{30}$	–	$\frac{4}{30}$
5	–	–	–	–	$\frac{2}{30}$	$\frac{2}{30}$
$\sum$	$\frac{3}{30}$	$\frac{14}{30}$	$\frac{7}{30}$	$\frac{3}{30}$	$\frac{3}{30}$	1

varajuće frekvencije, a u nutarnja polja tabl. 2. odgovarajuće relativne frekvencije uređenih parova iz danog niza statističkih podataka. Na donjoj margini upisane su vrijednosti zbroja frekvencija (relativnih frekvencija) dotičnog stupca, a na desnoj margini vrijednosti zbroja frekvencija (relativnih frekvencija) dotičnog retka. U donje krajnje desno polje talice upisana je vrijednost zbroja svih frekvencija (relativnih frekvencija).

Ako je općenito riječ o dva diskretna statistička obilježja X i Y, pri čemu obilježje X poprima vrijednosti iz diskretnog skupa A, a obilježje Y iz diskretnog skupa B, onda se prilikom simultanog promatranja (mjerenja, opažanja) tih obilježja kao rezultati dobivaju elementi skupa $A \times B$ (Kartezijev produkt skupova A i B). To znači da se za svaki uređeni par $(a, b) \in A \times B$ može govoriti o njegovoj frekvenciji $f(a, b)$ u nizu statističkih podataka $(x_1, y_1), \ldots, (x_n, y_n)$ o dvodimenzionalnome statističkom obilježju (X, Y). Također se govori i o relativnoj frekvenciji $p(a, b) = \frac{f(a, b)}{n}$ uređenog para (a, b) u danom nizu statističkih podataka.

Prema tome, za svaki niz statističkih podataka $(x_1, y_1), \ldots, (x_n, y_n)$ o diskretnome dvodimenzionalnom obilježju (X, Y) može se formirati pripadna tablica frekvencija koja se obično zove *kontingencijska tablica* (tabl. 3).

Tablica 3.

X \ Y	b_1	b_2	$\ldots$	b_k	$\ldots$	b_s	$\sum$
a_1	f_{11}	f_{12}	$\ldots$	f_{1k}	$\ldots$	f_{1s}	f_1
a_2	f_{21}	f_{22}	$\ldots$	f_{2k}	$\ldots$	f_{2s}	f_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_j	f_{j1}	f_{j2}	$\ldots$	f_{jk}	$\ldots$	f_{js}	f_j
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
a_r	f_{r1}	f_{r2}	$\ldots$	f_{rk}	$\ldots$	f_{rs}	f_r
$\sum$	g_1	g_2	$\ldots$	g_k	$\ldots$	g_s	n

Brojevi $a_1, \ldots, a_r$ elementi su skupa A, a brojevi $b_1, \ldots, b_s$ elementi su skupa B i u tabl. 3. obično su poredani po veličini tako da je $a_1 < a_2 < \ldots < a_r$ i $b_1 < b_2 < \ldots < b_s$.

Uređenom paru (a_j, b_k) pripada frekvencija

(1) $$f_{jk} = f(a_j, b_k), \quad j = 1, \ldots, r, \quad k = 1, \ldots, s,$$

odnosno relativna frekvencija

(2) $$p_{jk} = p(a_j, b_k) = \frac{1}{n} f_{jk}.$$

Očigledno je f_{jk} cijeli broj i vrijedi

(3) $$0 \leq f_{jk} \leq n, \quad \sum_{j=1}^{r} \sum_{k=1}^{s} f_{jk} = n,$$

(4) $$0 \leq p_{jk} \leq 1, \quad \sum_{j=1}^{r} \sum_{k=1}^{s} p_{jk} = 1.$$

Nadalje je

(5) $$f_j = \sum_{k=1}^{s} f_{jk}, \quad j = 1, \ldots, r,$$

(6) $$g_k = \sum_{j=1}^{r} f_{jk}, \quad k = 1, \ldots, s,$$

a definira se i

(7) $$p_j = \sum_{k=1}^{s} p_{jk} = \frac{1}{n} f_j, \quad j = 1, \ldots, r,$$

(8) $$q_k = \sum_{j=1}^{r} p_{jk} = \frac{1}{n} g_k, \quad k = 1, \ldots, s.$$

U 1. primjeru očigledno je $A = B = \{1,2,3,4,5\}$, tako da je $a_1 = b_1 = 1$, $a_2 = b_2 = 2$, $a_3 = b_3 = 3$, $a_4 = b_4 = 4$ i $a_5 = b_5 = 5$. Iz tabl. 1. vidi se, na primjer, da je $f_{32} = 5$, dok je $f_3 = 10$, a $g_2 = 14$. Iz tabl. 2. se, pak, vidi da je $p_{32} = \frac{5}{30} \approx 0{,}67$, $p_3 = \frac{10}{30} \approx 0{,}33$ i $q_2 = \frac{14}{30} \approx 0{,}47$. Veličina $f_{32} = 5$ označuje da u promatranom razredu ima 5 učenika koji imaju ocjenu 3 (dobar) iz matematike i ocjenu 2 (dovoljan) iz fizike. Veličina $f_3 = 10$ pokazuje da u tom razredu ima 10 učenika koji imaju ocjenu dobar (3) iz matematike, a veličina $g_2 = 14$ da u tom razredu ima 14 učenika s ocjenom dovoljan (2) iz fizike.

Općenito se može reći da broj f_j, definiran u (5), označuje frekvenciju vrijednosti a_j u nizu $x_1, \ldots, x_n$, dok broj g_k, definiran u (6), označuje frekvenciju vrijednosti b_k u nizu $y_1, \ldots, y_n$. Iz (7) i (8) se odmah razabire da su p_j i q_k odgovarajuće relativne frekvencije.

2. Dvodimenzionalna razdioba frekvencija

Tablicom 3, ili tablicom u kojoj bi umjesto frekvencija f_{jk} stajale relativne frekvencije p_{jk}, definirana je tzv. ***dvodimenzionalna razdioba frekvencija***, odnosno relativnih frekvencija, za dani niz statističkih podataka (x_1, y_1), ..., (x_n, y_n). Formulom (1) definirana je pripadna funkcija frekvencija, a formulom (2) pripadna funkcija relativnih frekvencija. Ovdje su to funkcije dviju varijabli, za razliku od onih u I. 2.

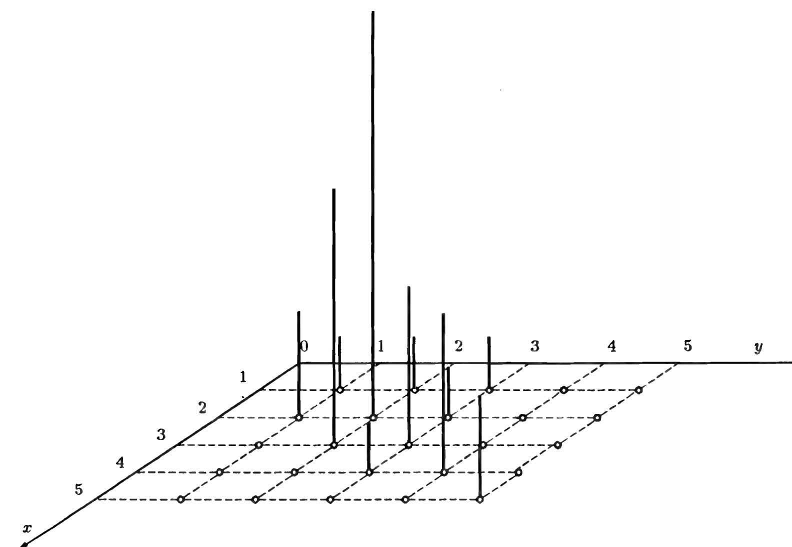
Odmah se uočava da se u vezi sa zadanom dvodimenzionalnom razdiobom frekvencija mogu promatrati i dvije obične (jednodimenzionalne) razdiobe frekvencija definirane formulama (5) i (6), odnosno (7) i (8). Formulama (5) i (7) određena je razdioba frekvencija i relativnih frekvencija za podatke o statističkom obilježju X, a formulama (6) i (8) za podatke o statističkom obilježju Y.

Prema tome, na marginama tabl. 3. mogu se promatrati dvije jednodimenzionalne razdiobe frekvencija, i to posebno za podatke o obilježju X i posebno za podatke o obilježju Y. Stoga se i govori o *marginalnim razdiobama frekvencija* za podatke o obilježjima X i Y, koje proizlaze iz zadane dvodimenzionalne razdiobe frekvencija za podatke o dvodimenzionalnome statističkom obilježju (X, Y).

Ako se, po analogiji s jednodimenzionalnim slučajem, žele geometrijski interpretirati statistički podaci o dvodimenzionalnome statističkom obilježju, onda se treba poslužiti prostornim (trodimenzionalnim) pravokutnim koordinatnim sustavom. Na apscisnu os postave se vrijednosti obilježja X, na ordinatnu os vrijednosti obilježja Y, a kao pripadne aplikate uzmu se odgovarajuće frekvencije, odnosno relativne frekvencije.

Podacima iz tabl. 1. odgovara sl. 17. Na sl. 17. zorno je prikazana dvodimenzionalna razdioba frekvencija za statističke podatke pri simultanom promatranju učeničkih ocjena iz matematike (obilježje X) i fizike (obilježje Y).

Slici 17. može se dati i mehanička interpretacija. Ako se, naime, brojčane vrijednosti frekvencija interpretiraju kao mase odgovarajućih točaka u ravnini x–y,



Slika 17. Prikaz podataka iz tabl. 1.

onda se može govoriti o određenoj *ravninskoj razdiobi mase* ukupne količine n. Ako su posrijedi relativne frekvencije, onda imamo ravninsku razdiobu mase ukupne količine 1. Time je uspostavljena određena analogija između dvodimenzionalne razdiobe frekvencija i ravninske razdiobe mase, što će omogućiti da se i drugim statističkim pojmovima u vezi s dvodimenzionalnim razdiobama daju mehaničke interpretacije.

3. Funkcije regresije

Kao što je već rečeno, glavni je zadatak pri sređivanju i obradi statističkih podataka da se otkriju određena svojstva i eventualne zakonitosti promatranih veličina X i Y. Odmah se uočava da se statistički podaci $x_1, \ldots, x_n$ o obilježju X mogu zasebno obrađivati svim onim metodama koje su opisane u I. i II. poglavlju, a isto tako se može reći i za podatke $y_1, \ldots, y_n$ o statističkom obilježju Y. To znači da se mogu definirati parametri

(9) $$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} \sum_{j=1}^{r} a_j f_j = \sum_{j=1}^{r} a_j p_j\,,$$

(10) $$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{k=1}^{s} b_k g_k = \sum_{k=1}^{s} b_k q_k\,,$$

(11)
$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{r}(a_j - \overline{x})^2 f_j = \sum_{j=1}^{r}(a_j - \overline{x})^2 p_j\,,$$

(12)
$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{k=1}^{s}(b_k - \overline{y})^2 g_k = \sum_{k=1}^{s}(b_k - \overline{y})^2 q_k\,.$$

Točka $(\overline{x},\overline{y})$ može se mehanički interpretirati kao težište odgovarajuće ravninske razdiobe mase, što onda opravdava i naziv ***središte dvodimenzionalne razdiobe frekvencija*** za točku $(\overline{x},\overline{y})$ u ravnini x–y danoga koordinatnog sustava.

Za statističke podatke iz 1. primjera dobiva se

$$\overline{x} = 2{,}70, \quad s_x^2 = 1{,}08\,, \quad s_x = 1{,}04\,,$$
$$\overline{y} = 2{,}63, \quad s_y^2 = 1{,}23\,, \quad s_y = 1{,}11\,.$$

Govori se da je prosječna ocjena učenika promatranog razreda iz matematike 2,70 uz standardnu devijaciju 1,04, dok je prosječna ocjena iz fizike 2,63 uz standardnu devijaciju 1,11. No pogled na tabl. 1. i 2. sugerira da je u njima sadržana i određena informacija o zavisnosti između veličina X i Y, a ne samo informacija o zasebnim svojstvima obilježja X i obilježja Y. Zato se i postavlja zadatak da se, uz pomoć danih statističkih podataka dobivenih pri simultanom promatranju (mjerenju) obilježja X i Y, istraži priroda međusobne ovisnosti veličina X i Y. To će se učiniti tako da se definiraju određeni parametri koji će kvantitativno izraziti tu ovisnost, te da se ona na određeni način geometrijski interpretira.

Svaki nutarnji redak i svaki nutarnji stupac, tablica 1, 2. i 3. može se shvatiti tako kao da je njime definirana određena jednodimenzionalna razdioba frekvencija. Tako, na primjer, izdvoji li se treći redak tabl. 1, dobiva se razdioba frekvencija prikazana u tabl. 4.

Tablica 4.

$X = 3$	Vrijednost obilježja Y	1	2	3	4	5
	Frekvencija	0	5	4	0	1

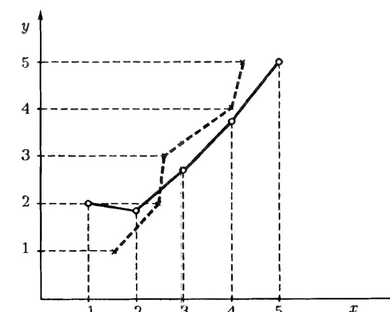
Tablica 4. pokazuje da od 10 učenika koji imaju ocjenu 3 iz matematike, ocjenu 1 iz fizike nema nitko, ocjenu 2 ima 5 učenika, ocjenu 3 ima 4 učenika i ocjenu 5 ima jedan učenik. Prosječna ocjena iz fizike za tih 10 učenika iznosi

$$\overline{y}(3) = \frac{1}{10}(1\cdot 0 + 2\cdot 5 + 3\cdot 4 + 4\cdot 0 + 5\cdot 1) = 2{,}7.$$

Na sličan način dobiva se

$$\overline{y}(1) = \frac{1}{3}(1\cdot 1 + 2\cdot 1 + 3\cdot 1) = 2,$$
$$\overline{y}(2) = \frac{1}{11}(1\cdot 1 + 2\cdot 8 + 3\cdot 1) = \frac{21}{11} \approx 1{,}91,$$
$$\overline{y}(4) = \frac{1}{4}(3\cdot 1 + 4\cdot 3) = 3{,}75,$$
$$\overline{y}(5) = \frac{1}{2}\cdot 5\cdot 2 = 5.$$

Sada se mogu na apscisnu os pravokutnoga koordinatnog sustava u ravnini nanijeti vrijednosti obilježja X, tj. ocjene iz matematike, a kao pripadne ordinate uzeti odgovarajuće prosječne ocjene iz fizike, što je prikazano na sl. 18.



Slika 18. Krivulje regresije za podatke iz tabl. 1.

Izlomljena puna crta na sl. 18. pokazuje kako se mijenja prosječna ocjena iz fizike u ovisnosti o ocjeni iz matematike.

Posve je razumljivo da X i Y mogu zamijeniti uloge, tj. da se svaki nutarnji stupac tabl. 1. tretira kao određena jednodimenzionalna razdioba frekvencija, koja pokazuje razdiobu ocjena iz matematike onih učenika koji imaju uočenu ocjenu iz fizike. Tako se iz tabl. 1. vidi da za 3 učenika koji imaju ocjenu 1 iz fizike, prosječna ocjena iz matematike iznosi

$$\overline{x}(1) = \frac{1}{3}(1\cdot 1 + 2\cdot 2) = \frac{5}{3} \approx 1{,}67.$$

Nadalje je

$$\overline{x}(2) = \frac{1}{14}(1\cdot 1 + 2\cdot 8 + 3\cdot 5) = \frac{32}{14} \approx 2{,}29,$$
$$\overline{x}(3) = \frac{1}{7}(1\cdot 1 + 2\cdot 1 + 3\cdot 4 + 4\cdot 1) = \frac{19}{7} \approx 2{,}71,$$

$$\overline{x}(4) = \frac{1}{3}\cdot 4\cdot 3 = 4,$$

$$\overline{x}(5) = \frac{1}{3}(3 \cdot 1 + 5 \cdot 2) = \frac{13}{3} \approx 4{,}33.$$

Na sl. 18. ucrtana je i izlomljena crta prikazana točkicama koja pokazuje kako se mijenja prosječna ocjena iz matematike u ovisnosti o ocjeni iz fizike.

Izlomljene crte na sl. 18. zovu se ***krivulje regresije*** za dani niz statističkih podataka o dvodimenzionalnom obilježju (X, Y) i one, na određeni način, pokazuju međusobnu zavisnost između ocjena iz matematike i ocjena iz fizike kod učenika promatranog razreda.

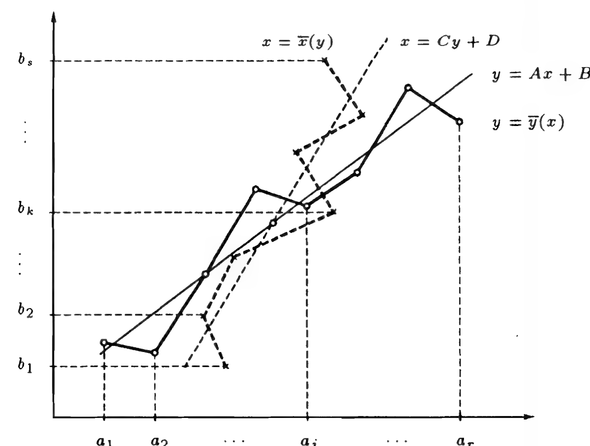
Općenito, polazeći od kontingencijske tablice (tabl. 3) za dani niz statističkih podataka o dvodimenzionalnom obilježju (X, Y), mogu se promatrati dvije familije jednodimenzionalnih razdioba frekvencija. Promatrajući nutarnje retke tabl. 3. uočava se r jednodimenzionalnih razdioba frekvencija, pri čemu j-ti redak određuje tzv. ***uvjetnu razdiobu frekvencija*** onih podataka o obilježju Y kod kojih obilježje X ima vrijednost a_j $(j = 1, \ldots, r)$. Vidi se da takvih podataka ukupno ima f_j. U toj razdiobi pripadni prosjek nazvat će se *uvjetni prosjek* i označiti sa $\overline{y}(a_j)$, te računati pomoću formule

$$\overline{y}(a_j) = \frac{1}{f_j}\sum_{k=1}^{s} b_k f_{jk}, \quad j = 1, \ldots, r. \tag{13}$$

Promatrajući, pak nutarnje stupce tabl. 3, uočava se s jednodimenzionalnih razdioba frekvencija, pri čemu k-ti $(k = 1, \ldots, s)$ redak određuje uvjetnu razdiobu frekvencija onih podataka o obilježju X, ukupno ih ima g_k, kod kojih obilježje Y ima vrijednost b_k. U to razdiobi uvjetni prosjek označuje se sa $\overline{x}(b_k)$ i računa pomoću formule

$$\overline{x}(b_k) = \frac{1}{g_k}\sum_{j=1}^{r} a_j f_{jk}, \quad k = 1, \ldots, s. \tag{14}$$

Formulama (13) i (14) definirane su funkcije $a_j \mapsto \overline{y}(a_j)$ i $b_k \mapsto \overline{x}(b_k)$ koje se zovu *funkcije regresije* danog niza statističkih podataka o dvodimenzionalnom obilježju (X, Y). Funkcija $a_j \mapsto \overline{y}(a_j)$, $j = 1, \ldots, r$, pokazuje ovisnost prosjeka uvjetne razdiobe frekvencija u j-tom retku tabl. 3. o vrijednosti a_j obilježja X, a funkcija $b_k \mapsto \overline{x}(b_k)$, $k = 1, \ldots, s$, pokazuje ovisnost prosjeka uvjetne razdiobe frekvencija u k-tom stupcu tabl. 3. o vrijednosti b_k obilježja Y. Grafički prikaz tih funkcija u pravokutnome koordinatnom sustavu redovito će imati oblik dviju izlomljenih crta (sl. 19).



Slika 19. Krivulje regresije i pravci regresije

4. Pravci regresije

Prirodno se nameće ideja da se izlomljene crte, koje zorno prikazuju funkcije regresije, zamjene pravcima (sl. 19). Parametri A i B, odnosno C i D, odredit će se *metodom najmanjih kvadrata* s odgovarajućim težinskim faktorima. To znači da parametre A i B treba odrediti tako da zbroj

$$S = \sum_{j=1}^{r}\sum_{k=1}^{s}[Aa_j + B - b_k]^2 f_{jk} \tag{15}$$

bude minimalan. To će se postići rješavanjem sustava jednadžbi

$$\frac{\partial S}{\partial A} = 0, \qquad \frac{\partial S}{\partial B} = 0.$$

Nakon provedenoga parcijalnog deriviranja i sređivanja dobiva se sustav

$$\begin{cases} A\sum_{j=1}^{r}\sum_{k=1}^{s} a_j^2 f_{jk} + B\sum_{j=1}^{r}\sum_{k=1}^{s} a_j f_{jk} = \sum_{j=1}^{r}\sum_{k=1}^{s} a_j b_k f_{jk} \\ A\sum_{j=1}^{r}\sum_{k=1}^{s} a_j f_{jk} + B\sum_{j=1}^{r}\sum_{k=1}^{s} f_{jk} = \sum_{j=1}^{r}\sum_{k=1}^{s} b_k f_{jk}. \end{cases} \tag{16}$$

Imajući na umu formule (3), (5), (6), (9), (10) i (11) lako se uočava da je

(17) $$\sum_{j=1}^{r}\sum_{k=1}^{s} f_{jk} = n,$$

(18) $$\sum_{j=1}^{r}\sum_{k=1}^{s} a_j f_{jk} = n\overline{x},$$

(19) $$\sum_{j=1}^{r}\sum_{k=1}^{s} b_k f_{jk} = n\overline{y},$$

(20) $$\sum_{j=1}^{r}\sum_{k=1}^{s} a_j^2 f_{jk} = n(s_x^2 + \overline{x}^2).$$

Uvede li se još oznaka

(21) $$s_{xy} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s}(a_j - \overline{x})(b_k - \overline{y})f_{jk} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} a_j b_k f_{jk} - \overline{x}\,\overline{y},$$

odmah se vidi da je

(22) $$\sum_{j=1}^{r}\sum_{k=1}^{s} a_j b_k f_{jk} = n(s_{xy} + \overline{x}\,\overline{y}).$$

Uzme li se u obzir (17), (18), (19), (20) i (22), sustav (16) postaje

$$\begin{aligned} &A\, n(s_x^2 + \overline{x}^2) + Bn\overline{x} = n(s_{xy} + \overline{x}\,\overline{y}) \\ &A\, n\overline{x} + Bn = n\overline{y}\,, \end{aligned}$$

odnosno

(23) $$\begin{cases} A(s_x^2 + \overline{x}^2) + B\overline{x} = s_{xy} + \overline{x}\,\overline{y} \\ A\overline{x} + B = \overline{y}\,. \end{cases}$$

Rješavanjem sustava (23) po nepoznanicama A i B dobiva se

(24) $$A = \frac{s_{xy}}{s_x^2}, \qquad B = \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x},$$

što, dakako, ima smisla samo za $s_x > 0$.

Prema tome, jednadžba pravca koji u smislu metode najmanjih kvadrata najbolje aproksimira krivulju regresije $y = \overline{y}(a_j)$, glasi

$$y = Ax + B = \frac{s_{xy}}{s_x^2}x + \overline{y} - \frac{s_{xy}}{s_x^2}\overline{x},$$

odnosno

(25) $$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x}).$$

Ako se vrijednosti za A i B iz (24) uvrste u (15), nakon sređivanja dobiva se

(26) $$S = ns_y^2\left(1 - \frac{s_{xy}^2}{s_x^2 s_y^2}\right).$$

Ako je i $s_y > 0$, onda se može uvesti oznaka

(27) $$r = \frac{s_{xy}}{s_x s_y}\,,$$

pa (26) postaje

(28) $$S = ns_y^2(1 - r^2).$$

Na posve analogan način mogu se odrediti parametri C i D tako da zbroj

(29) $$T = \sum_{j=1}^{r}\sum_{k=1}^{s}(Cb_k + D - a_j)^2 f_{jk}$$

bude minimalan. Dobiva se

(30) $$C = \frac{s_{xy}}{s_y^2}, \qquad D = \overline{x} - \frac{s_{xy}}{s_y^2}\overline{y},$$

tako da jednadžba pravca, koji u smislu metode najmanjih kvadrata najbolje aproksimira krivulju regresije $x = \overline{x}(b_k)$, glasi

(31) $$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y}).$$

Ako se vrijednosti za C i D iz (30) uvrste u (29), nakon sređivanja dobiva se

(32) $$T = ns_x^2(1 - r^2).$$

Iz (25) i (31) razabire se da dobiveni pravci, koji se zovu ***pravci regresije***, prolaze točkom $(\overline{x}, \overline{y})$, što pokazuje da se oni sijeku u središtu dane dvodimenzionalne razdiobe frekvencija. Za kut φ koji čine ti pravci očigledno vrijedi

(33) $$\operatorname{tg}\varphi = \frac{1 - r^2}{r}\,\frac{s_x s_y}{s_x^2 + s_y^2}.$$

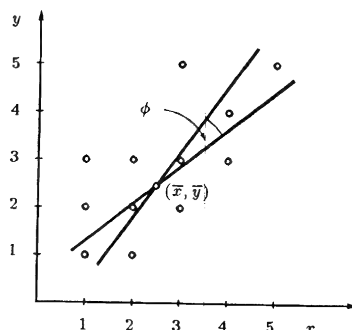
Za podatke iz 1. primjera izračunali smo već da je $\overline{x} = 2{,}70$, $\overline{y} = 2{,}63$, $s_x = 1{,}04$ i $s_y = 1{,}11$, a primjenom formule (21) dobiva se

$$\begin{aligned} s_{xy} = \frac{1}{30}(&1\cdot 1\cdot 1 + 1\cdot 2\cdot 1 + 1\cdot 3\cdot 1 + 2\cdot 1\cdot 2 + 2\cdot 2\cdot 8 + 2\cdot 3\cdot 1 + 3\cdot 2\cdot 5 + \\ &+3\cdot 3\cdot 4 + 3\cdot 5\cdot 1 + 4\cdot 3\cdot 1 + 4\cdot 4\cdot 3 + 5\cdot 5\cdot 2) - 2{,}70\cdot 2{,}63 \approx 0{,}87, \end{aligned}$$

pa se odmah mogu napisati jednadžbe pravaca regresije

$$y - 2{,}63 = 0{,}80(x - 2{,}70)$$
$$x - 2{,}70 = 0{,}71(y - 2{,}63).$$

Kako je $r = \dfrac{0{,}87}{1{,}04 \cdot 1{,}11} = 0{,}75$, iz (33) se dobiva $\operatorname{tg}\varphi = 0{,}576 \cdot \dfrac{1{,}1544}{2{,}3137} = 0{,}2874$, iz čega proizlazi da je kut između pravaca regresije $\varphi \approx 16°$.



Slika 20. Pravci regresije za podatke iz 1. primjera

5. Koeficijent korelacije

Budući da su S i T nenegativne veličine, što se vidi iz definicijskih formula (15) i (29), iz (28) i (32) proizlazi da je $1 - r^2 \geq 0$, odnosno

(34) $$r^2 \leq 1, \qquad -1 \leq r \leq 1.$$

Ako je $r^2 = 1$, onda se iz (33) razabire da je $\varphi = 0$, ili $\varphi = \pi$, što znači da se pravci regresije međusobno poklapaju. Nadalje se iz (28) i (32) vidi da je tada $S = T = 0$, što znači da se dani statistički podaci $(x_1, y_1), \ldots, (x_n, y_n)$, odnosno odgovarajuće točke, nalaze na tom zajedničkom pravcu koji ima jednadžbu

(35) $$y - \overline{y} = \frac{s_y}{s_x}(x - \overline{x}).$$

To, pak, znači da je

(36) $$y_i = \frac{s_y}{s_x}(x_i - \overline{x}) + \overline{y}, \quad i = 1, \ldots, n,$$

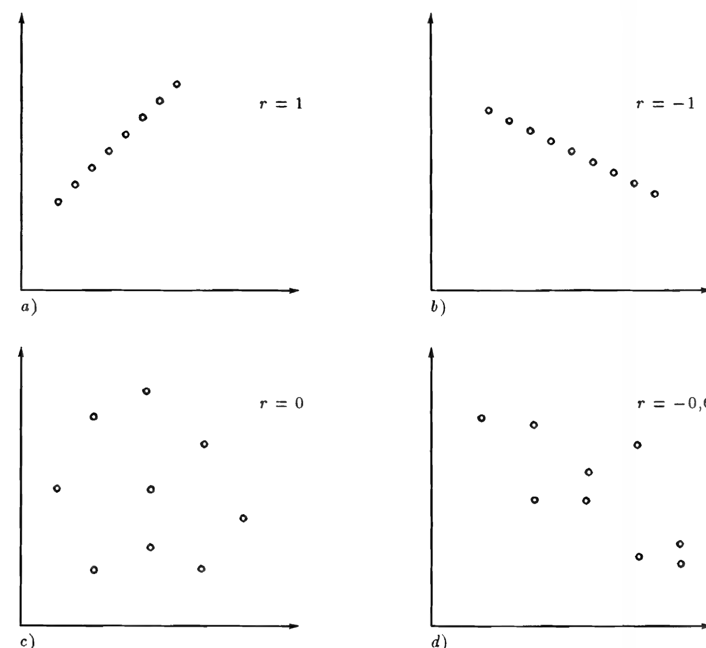
tj. podaci o obilježju Y funkcijski su ovisni o podacima o obilježju X, preko afine funkcije izražene formulom (36).

Ako je $r = 0$, onda je i $s_{xy} = 0$, što se vidi iz (27), dok se iz (25) vidi da je prvi pravac regresije usporedan s apscisnom osi, a iz (31) se vidi da je drugi pravac regresije usporedan s ordinatnom osi. To znači da su zadani statistički podaci tako raspoređeni u koordinatnom sustavu da je linearna aproksimacija funkcije regresije $a_j \mapsto \overline{y}(a_j)$ neovisna o a_j $(j = 1, \ldots, r)$, a također i linearna aproksimacija funkcije regresije $b_k \mapsto \overline{x}(b_k)$ neovisna je o b_k $(k = 1, \ldots, s)$. U tom se slučaju govori da su x_i i y_i $(i = 1, \ldots, n)$ ***nekorelirani*** statistički podaci.

Općenito se parametar r, definiran u (27), zove ***koeficijent korelacije*** i ako je $r < 0{,}5$ kaže se da su x_i i y_i *slabo korelirani* podaci, a ako je $r \geq 0{,}5$ kaže se da je korelacija značajna.

Za $r > 0$ govori se o *pozitivnoj korelaciji*, koja upućuje na činjenicu da se s porastom vrijednosti iksova (ipsilona) i vrijednosti ipsilona (iksova) u prosjeku povećavaju. Za $r < 0$ govori se o *negativnoj korelaciji*, što upućuje na prosječno smanjivanje vrijednosti ipsilona (iksova) pri povećanju vrijednosti iksova (ipsilona).

Za podatke iz 1. primjera izračunali smo $r = 0{,}75$, pa se može reći da su u promatranom razredu ocjene učenika iz matematike značajno pozitivno korelirane s ocjenama učenika iz fizike.



Slika 21. Skica rasporeda podataka za različite vrijednosti koeficijenata korelacije

Iz iznesenoga se vidi da koeficijent korelacije, kao određeni parametar koji se odnosi na statističke podatke o dvodimenzionalnom obilježju (X, Y), pokazuje stupanj afine funkcijske povezanosti među podacima o obilježju X i podacima o obilježju Y. Za $r = 1$, ili $r = -1$, postoji potpuna afina (ponekad se govori i linearna) veza između iksova i ipsilona, dok za $r = 0$ nema govora o linearnoj povezanosti iksova i ipsilona. No, time još nije utvrđeno da ne postoji nikakva statistička povezanost između iksova i ipsilona.

6. Statistička zavisnost

Pri izučavanju problema zavisnosti i nezavisnosti statističkih podataka o diskretnom, dvodimenzionalnom obilježju (X, Y) čini se razumnim poći od kontingencijske tablice (v. tabl. 3). Pokazalo se, naime, zanimljivim izučiti one kontingencijske tablice koje zadovoljavaju uvjet:

(37) $$nf_{jk} = f_j g_k, \quad j = 1, \ldots, r, \quad k = 1, \ldots, s.$$

Ako kontingencijska tablica ispunjava uvjet (37), onda se uvjetne razdiobe frekvencija u recima tablice eventualno međusobno razlikuju samo u određenom koeficijentu proporcionalnosti. Analogna konstatacija vrijedi i za stupce takve kontingencijske tablice.

Prema tome, u kontingencijskoj tablici koja ispunjava uvjet (37) sve uvjetne razdiobe frekvencije podataka o obilježju Y na određeni su način međusobno slične, pa se može reći da vrijednost obilježja X bitno ne utječe na uvjetnu razdiobu frekvencija podataka o obilježju Y. Također se može reći da vrijednost obilježja Y bitno ne utječe na uvjetnu razdiobu frekvencija podataka o obilježju X.

Razumno je, stoga, reći da su podaci o obilježju X i podaci o obilježju Y, prikazani kontingencijskom tablicom koja ispunjava uvjet (37), ***statistički nezavisni***.

Da bi se dobio pokazatelj odstupanja od statističke nazavisnosti, prirodno je promatrati razlike $nf_{jk} - f_j g_k$, $j = 1, \ldots, r$, $k = 1, \ldots, s$ i pomoću njih definirati parametar

(38) $$f^2 = \frac{1}{n^2} \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(nf_{jk} - f_j g_k)^2}{f_j g_k} = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{f_{jk}^2}{f_j g_k} - 1,$$

koji će globalno pokazivati ***odstupanje od statističke nezavisnosti*** u danoj kontingencijskoj tablici.

Iz (38) se razabire da vrijedi

(39) $$0 \leq f^2 \leq \min\{r, s\} - 1$$

i $f = 0$ postiže se onda i samo onda kada je ispunjen uvjet (37), tj. kada je riječ o statističkoj nezavisnosti podataka. Nadalje, $f^2 = \min\{r, s\} - 1$ postiže se, pri $r \geq s$, onda i samo onda kada svaki redak kontingencijske tablice sadrži samo jednu, od nule različitu, vrijednost, a pri $r \leq s$ onda i samo onda kada svaki stupac kontingencijske tablice sadrži samo jednu, od nule različitu, frekvenciju. To, pak, znači da parametar f^2 poprima maksimalnu moguću vrijednost onda i samo onda kada su iksovi i ipsiloni povezani funkcionalnom zavisnošću, tj. svakoj vrijednosti a_j $(j = 1, \ldots, r)$ obilježja X pridružena je jedna i samo jedna vrijednost obilježja Y (slučaj $r \geq s$), ili je, pak, svakoj vrijednosti b_k $(k = 1, \ldots, s)$ obilježja Y pridružena jedna i samo jedna vrijednost obilježja X (slučaj $r \leq s$).

Za podatke iz 1. primjera dobiva se

$$f^2 = \frac{1^2}{3 \cdot 3} + \frac{1^2}{3 \cdot 14} + \frac{1^2}{3 \cdot 7} + \frac{2^2}{11 \cdot 3} + \frac{8^2}{11 \cdot 14} + \frac{1^2}{11 \cdot 7} + \frac{5^2}{10 \cdot 14} + \frac{4^2}{10 \cdot 7} + \frac{1^2}{10 \cdot 3} + \\ + \frac{1^2}{4 \cdot 7} + \frac{3^2}{4 \cdot 3} + \frac{2^2}{2 \cdot 3} - 1 = 1{,}625.$$

Maksimalna vrijednost za f^2, u ovom primjeru $(r = s = 5)$, iznosi 4, pa se može zaključiti da ocjena iz matematike i ocjena iz fizike u promatranom razredu nisu statistički nezavisne veličine, ali su i vrlo daleko od međusobne funkcijske zavisnosti.

Veličina

(40) $$o = \frac{f^2}{\min\{r, s\} - 1}$$

pokazuje ***stupanj statističke zavisnosti*** iksova i ipsilona u danom nizu statističkih podataka o dvodimenzionalnom obilježju (X, Y).

Za podatke iz 1. primjera dobiva se stupanj statističke zavisnosti $o = \frac{1{,}625}{4} \approx$ $\approx 0{,}41$, pa se može reći da između ocjena iz matematike i ocjena iz fizike u promatranom razredu postoji zavisnost od približno 41 %.

Zanimljivo je primijetiti da u definiciji parametra f^2 ne sudjeluju vrijednosti obilježja X i Y, već samo frekvencije iz kontingencijske tablice. To omogućuje da se parametar f^2, a također i stupanj statističke zavisnosti o, primijeni kao pokazatelj statističke zavisnosti i u slučaju nenumeričkih statističkih obilježja.

2. primjer

Istraživana je veza između pojave povišenoga krvnog tlaka i pušenja, tako da je ispitano 180 osoba i rezultati su prikazani u tabl. 5.

Tablica 5.

	Nepušač	Blagi pušač	Teški pušač	$\sum$
Normalni tlak	48	26	19	93
Povišeni tlak	21	36	30	87
$\sum$	69	62	49	180

Izračuna li se, primjenom formule (38), vrijednost parametra f^2, dobiva se $f^2 = 0{,}08$. Kako maksimalna moguća vrijednost za f^2 u ovom primjeru $(r = 2, s = 3)$ iznosi $\min\{2, 3\} - 1 = 1$, može se zaključiti da je u promatranoj skupini osoba povišeni krvni tlak vrlo slabo povezan s pušačkim statusom osobe. Stupanj statističke zavisnosti iznosi samo 8 %.

7. Kontinuirana statistička obilježja

Sva dosadašnja razmatranja odnosila su se na diskretna obilježja X i Y. Ako se niz statističkih podataka $(x_1, y_1), \ldots, (x_n, y_n)$ odnosi na kontinuirana obilježja X i Y, onda se za oblikovanje kontingencijske tablice najprije mora izvršiti grupiranje podataka o obilježju X u, recimo, r razreda i grupiranje podataka o obilježju Y u, recimo, s razreda. To se radi na isti način kao što je opisano za jednodimenzionalne razdiobe frekvencija u I. poglavlju. Nakon toga određuje se broj f_{jk} (frekvencija) onih uređenih parova u nizu $(x_1, y_1), \ldots, (x_n, y_n)$ kod kojih prvi član uređenog para pripada j-tom razredu iksova, a drugi član k-tom razredu ipsilona. Kontingencijska će tablica za podatke o dva kontinuirana statistička obilježja X i Y obično imati oblik prikazan tablicom 6.

Ako su $a_0 < a_1 < \ldots < a_r$ rubovi razreda za iksove, a $b_0 < b_1 < \ldots < b_s$ rubovi razreda za ipsilone, onda su:

$$\bar{a}_j = \frac{1}{2}(a_{j-1} + a_j), \quad j = 1, \ldots, r,$$

$$\bar{b}_k = \frac{1}{2}(b_{k-1} + b_k), \quad k = 1, \ldots, s$$

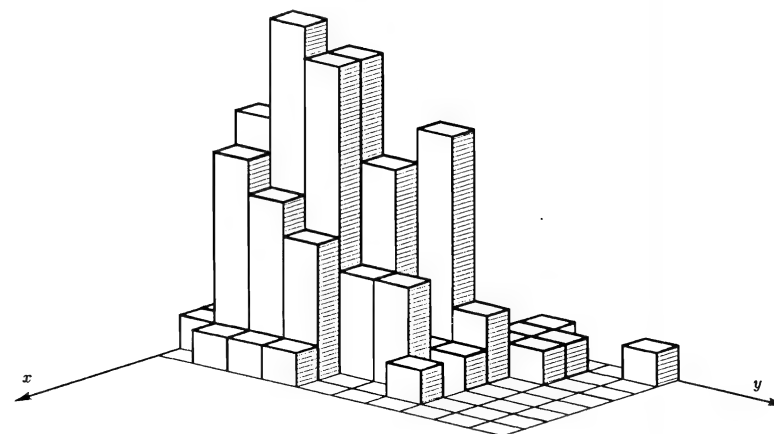
odgovarajuće sredine razreda.

Očigledno je da za veličine f_{jk}, f_j i g_k ($j = 1, \ldots, r$, $k = 1, \ldots, s$) iz tabl. 6. također vrijede formule (3), (5) i (6). Ako se stavi $p_{jk} = \frac{1}{n} f_{jk}$, tj. ako se radi s relativnim frekvencijama p_{jk}, onda vrijede i formule (4), (7) i (8).

Zorna interpretacija dvodimenzionalne razdiobe frekvencija za podatke o kontinuiranima statističkim obilježjima X i Y može se izvesti slično kao i za diskretna obilježja (sl. 17). Sada će se, samo, umjesto točke (a_j, b_k), frekvencijom f_{jk} "opteretiti" pravokutnik određen j-tim razredom iksova i k-tim razredom ipsilona, tako da se iznad njega ucrta kvadar visine f_{jk} (sl. 22).

Tablica 6.

Redni broj	Y	1	2	...	k	...	s	
X	Sredina razreda	$\bar{b}_1$	$\bar{b}_2$	...	$\bar{b}_k$	...	$\bar{b}_s$	$\sum$
1	$\bar{a}_1$	f_{11}	f_{12}	...	f_{1k}	...	f_{1s}	f_1
2	$\bar{a}_2$	f_{21}	f_{22}	...	f_{2k}	...	f_{2s}	f_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
j	$\bar{a}_j$	f_{j1}	f_{j2}	...	f_{jk}	...	f_{js}	f_j
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
r	$\bar{a}_r$	f_{r1}	f_{r2}	...	f_{rk}	...	f_{rs}	f_r
	$\sum$	g_1	g_2	...	g_k	...	g_s	n



Slika 22. Skica dvodimenzionalne razdiobe frekvencija za podatke o kontinuiranim obilježjima

Iz formula (9), (10), (11) i (12) vidljivo je da se veličine $\bar{x}$, $\bar{y}$, s_x^2 i s_y^2 mogu izračunati i bez uporabe frekvencija, što znači da te veličine ne ovise o izvršenom grupiranju podataka u razrede. Iz (21) se vidi da se i veličina s_{xy} može izraziti bez uporabe frekvencija, tj. može se pisati

$$(41) \qquad s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Pogledaju li se jednadžbe (25) i (31) pravaca regresije i definicijska formula (27) za koeficijent korelacije, odmah se vidi da su ti pojmovi korektno definirani i za podatke o kontinuiranim statističkim obilježjima X i Y, a to znači da se i za njih može govoriti o koreliranosti i nekoreliranosti podataka, te o značenju manje ili veće koreliranosti u danom nizu podataka.

Međutim, parametar f^2, definiran u (38), bitno ovisi o kontingencijskoj tablici, tako da će za podatke o kontinuiranim obilježjima X i Y vrijednost f^2 ovisiti i o primijenjenom načinu grupiranja podataka u razrede. Stoga će stupanj statističke ovisnosti o, definiran formulom (40), osim o danim podacima, ovisiti i o izvedenom grupiranju podataka u razrede, što je nepoželjno svojstvo jednoga takvog parametra i zato se izbjegava njegova primjena u praksi.

Proračun najvažnijih parametara, kao što su $\bar{x}$, $\bar{y}$, s_x^2, s_y^2 i s_{xy}, može se i u slučaju kontinuiranih obilježja izvesti na temelju sređenih podataka u odgovarajućoj kontingencijskoj tablici (tabl. 6) koja je načinjena primjenom grupiranja podataka u razrede. No, tako dobivene vrijednosti samo su približno jednake točnim vrijednostima tih parametara i zato se piše

$$(42) \qquad \bar{x} \approx \frac{1}{n} \sum_{j=1}^{r} a_j f_j = \frac{1}{n} \sum_{j=1}^{r} \sum_{k=1}^{s} a_j f_{jk},$$

$$\overline{y} \approx \frac{1}{n}\sum_{k=1}^{s} b_k g_k = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} b_k f_{jk}, \tag{43}$$

$$s_x^2 \approx \frac{1}{n}\sum_{j=1}^{r}(a_j - \overline{x})^2 f_j = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} a_j^2 f_{jk} - \overline{x}^2, \tag{44}$$

$$s_y^2 \approx \frac{1}{n}\sum_{k=1}^{s}(b_k - \overline{y})^2 g_k = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} b_k^2 f_{jk} - \overline{y}^2, \tag{45}$$

$$s_{xy} \approx \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s}(a_j - \overline{x})(b_k - \overline{y}) f_{jk} = \frac{1}{n}\sum_{j=1}^{r}\sum_{k=1}^{s} a_j b_k f_{jk} - \overline{x}\,\overline{y}. \tag{46}$$

Primjedba

Primjena elektroničkih računala (kompjutora) i odgovarajućih programskih paketa bitno olakšava i ubrzava rješavanje praktičnih zadataka u vezi s dvodimenzionalnim statističkim obilježjima, jer omogućuje brzo, pregledno i prilično točno dobivanje traženih numeričkih rezultata, a također i grafičke prikaze relevantnih pojmova (dvodimenzionalne razdiobe, pravaca regresije i sl.).

Zadaci

1. Dvije igraće kocke bačene su $n = 40$ puta. Kocke se međusobno razlikuju i prvi član (X) uređenog para odnosi se na prvu kocku, a drugi član (Y) na drugu kocku. Dobiveni su ovi rezultati:

x	6	4	6	5	6	1	3	1	3	2	2	5	3	4	6	4	1	2	3	4	1	5	3	4	1	6	6	4	2	3	6	2	4	2
y	1	4	2	6	3	1	3	5	3	2	2	1	6	3	3	4	3	1	4	5	1	4	4	5	2	1	3	3	5	2	3	6	5	4

x	5	2	3	4	2	3
y	3	1	1	4	3	2

a) Načinite odgovarajuću kontingencijsku tablicu.

b) Nađite marginalne razdiobe frekvencija.

c) Izračunajte $\overline{x}$, $\overline{y}$, s_x^2, s_y^2 i r.

d) Nađite pravce regresije.

e) Skicirajte u koordinatnom sustavu pripadnu dvodimenzionalnu razdiobu, funkcije regresije i pravce regresije.

2. Istraživan je odnos između ocjene (X) iz matematike u završnom razredu srednje škole i ocjene (Y) na fakultetskom ispitu iz matematike. Promotreno je $n = 172$ studenata i rezultati su prikazani ovom tablicom:

x \ y	1	2	3	4	5
2	3	2	–	1	–
3	6	28	10	–	1
4	2	1	10	48	10
5	1	–	7	12	30

Izračunajte koeficijente korelacije i kut između pravaca regresije.

3. Za $n = 30$ zadanih statističkih podataka

x	y	x	y	x	y
1,5	–0,1	1,2	–0,3	1,4	–0,1
0,6	0,7	0,9	1,0	1,5	0,5
1,4	–0,2	2,5	0	1,1	0,2
3,5	0,1	1,4	0,1	0,9	–1,2
2,3	0	4,7	–0,4	1,8	–0,2
0,8	0,3	1,1	–0,2	2,3	0,4
1,0	0,3	3,9	0,1	2,5	0,5
1,3	0,5	3,8	–0,2	3,0	–0,6
3,0	–0,4	0,9	0,8	1,3	0,5
0,8	0,8	4,1	0,8	1,5	1,1

primjenom grupiranja u razrede:

a) načinite kontingencijsku tablicu,

b) izračunajte koeficijent korelacije,

c) nađite stupanj statističke zavisnosti.

4. Za niz podataka

x	8,3	7,5	4,5	7,5	7,6	6,9	4,5	8,1	8,4	7,4
y	8,0	7,5	7,5	9,5	10	8,8	4,5	8,1	9,2	7,2

izračunajte koeficijent korelacije.

5. Za podatke svrstane u kontingencijsku tablicu

X \ Y	A	B	C	D
a	4	2	–	–
b	1	11	6	1
c	–	7	10	2
d	–	–	3	3

izračunajte stupanj statističke zavisnosti.

6. Želi se istražiti statistička zavisnost između susjednih slova u tekstovima hrvatskog jezika. Uzmite iz proizvoljnog teksta $n = 200$ uređenih parova susjednih slova i načinite pripadnu kontingencijsku tablicu, tako da obilježje X označuje prvo, a obilježje Y drugo slovo uređenog para. Na temelju tako dobivene kontingencijske tablice izračunajte stupanj statističke zavisnosti prvog i drugog slova uređenog para u danom nizu podataka.

7. Dokažite relacije (18)–(20).

8. Dokažite da vrijedi

$$\sum_{j=1}^{r}\sum_{k=1}^{s}(a_j-\overline{x})(b_k-\overline{y})f_{jk}=\sum_{j=1}^{r}\sum_{k=1}^{s}a_jb_kf_{jk}-n\overline{x}\,\overline{y}.$$

(relacija (21))

9. Dokažite da je formulama (24) dano rješenje sustava (23).

10. Izvedite formulu (26).

11. Izvedite formulu (33).

12. Dokažite relaciju (39).

Pregled važnijih pojmova i formula deskriptivne statistike

Pojam	Oznaka	Definicijska formula
frekvencija	f_j	broj pojavljivanja j-tog ($j=1,\ldots,r$) događaja u nizu od n opažanja
relativna frekvencija	p_j	$=\frac{1}{n}f_j$
funkcija kumulativnih relativnih frekvencija	$F(x)$	$=\sum_{a_j\le x}p_j$
prosjek	$\overline{x}$	$=\frac{1}{n}\sum_{i=1}^{n}x_i$
varijanca	s_0^2	$=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$
standardna devijacija	s_0	$=\sqrt{s_0^2}$
raspon	d	$=\max\{x_1,\ldots,x_n\}-\min\{x_1,\ldots,x_n\}$
centralni moment k-tog reda	m_k	$=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^k,\quad k=0,1,\ldots$
koeficijent asimetrije	K	$=\frac{m_3}{s_0^3}$
koeficijent spljoštenosti	E	$=\frac{m_4}{s_0^4}-3$
središte dvodimenzionalne razdiobe	$(\overline{x},\overline{y})$	$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$
korelacijski moment	s_{xy}	$=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$
koeficijent korelacije	r	$=\frac{s_{xy}}{s_xs_y}$
kut između pravaca regresije	φ	$=\operatorname{arctg}\left(\frac{1-r^2}{r}\frac{s_xs_y}{s_x^2+s_y^2}\right)$
odstupanje od statističke nezavisnosti	f^2	$=\sum_{j=1}^{r}\sum_{k=1}^{s}\frac{f_{jk}^2}{f_jg_k}-1$
stupanj statističke zavisnosti	o	$=\frac{f^2}{\min\{r,s\}-1}$

DRUGI DIO

MATEMATIČKA TEORIJA STATISTIČKIH FENOMENA

Temeljna je pretpostavka za stvaranje teorije o statističkim fenomenima koji se očituju pri proučavanju i analizi statističkih podataka da su izmjereni, odnosno opaženi, brojčani podaci posljedica postojanja tzv. *statističkih zakonitosti*, koje su i inače karakteristične za slučajne pojave.

Slučajnost rezultata mjerenja može biti posljedica prirode promatrane pojave, može biti prouzročena nedovoljnom preciznošću instrumenata za mjerenje, može biti uzrokovana subjektivnim faktorom (mjeračeva nesavršenost i sl.), a također može biti i rezultat svega zajedno. Iskustvo je, međutim, pokazalo da i uz prisutnost različitih vrsta slučajnosti postoje određene zakonitosti u globalnom ponašanju rezultata mjerenja, koje se mogu i matematički opisati.

Statističke zakonitosti očituju se u situacijama kada je broj n mjerenja "dovoljno velik", jer se tada relativne frekvencije stabiliziraju oko fiksiranih brojeva – *vjerojatnosti*. Pri izgradnji matematičkih modela statističkih fenomena korisno je zamisliti da je broj mjerenja (opažanja) beskonačno velik, jer će se tada, umjesto relativne frekvencije kao određene empirijske veličine ovisne o broju mjerenja n, u modelu pojaviti apstraktno-matematički pojam vjerojatnosti događaja. Stoga će se u statističkoj teoriji operirati s *razdiobama vjerojatnosti*, umjesto s razdiobama relativnih frekvencija, što je činjeno u deskriptivnoj statistici. Umjesto o statističkom obilježju X, u matematičkoj teoriji statističkih fenomena (*teorija slučajnih varijabli*) govori se o *slučajnoj varijabli* X, kojoj pripada odgovarajuća razdioba (*distribucija*) vjerojatnosti, što je posve teorijski pojam.

Teorijski model za simultano promatranje više statističkih obilježja $X_1, \ldots, X_k$ ($k \in \mathbf{N}$) razvijen je u okviru *teorije slučajnih vektora*, gdje je temeljni pojam *k-dimenzionalni slučajni vektor* $(X_1, \ldots, X_k)$, čije su komponente $X_1, \ldots, X_k$ slučajne varijable i kojemu pripada određena *k-dimenzionalna razdioba vjerojatnosti*.

Teorija slučajnih varijabli i slučajnih vektora dio je opsežnije matematičke discipline – *teorije vjerojatnosti*, koja je glavni teorijski oslonac za razvijanje teorije statističkih zakonitosti.

Budući da je osnovna svrha ove knjige da se iznesu temeljni pojmovi i metode matematičke statistike, te da se prikažu tipične primjene u istraživanjima i praksi u različitim strukama, u ovom dijelu knjige prikazat će se samo nužni pojmovi i navesti glavni rezultati teorije slučajnih varijabli i slučajnih vektora, bez dubljeg ulaženja u samu teoriju.

IV. Teorijska interpretacija jednodimenzionalnih statističkih obilježja

1. Razdioba vjerojatnosti

Razdioba vjerojatnosti slučajne varijable X karakterizirana je svojom ***funkcijom razdiobe vjerojatnosti*** (f.r.v.) definiranom formulom

$$(1) \qquad F(x) = \mathrm{P}(X \leq x), \quad x \in \mathbf{R}.$$

Formula (1) kazuje da je $F(x)$ realan broj koji označuje vjerojatnost da se prilikom mjerenja slučajne varijable (s.v.) X dobije vrijednost koja ne premašuje realni broj x.

Budući da je f.r.v. teorijski analogon funkcije kumulativnih relativnih frekvencija (v. formulu (5) i sl. 3. u I.2), očigledno vrijedi

$$(2) \qquad F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \quad F(\infty) = \lim_{x \to \infty} F(x) = 1,$$

$$(3) \qquad x_1 < x_2 \Longrightarrow F(x_1) \leq F(x_2), \quad x_1, x_2 \in \mathbf{R}.$$

Iz (1)-(3) vidi se da je $x \mapsto F(x)$ neopadajuća realna funkcija realne varijable, koja može poprimati vrijednosti iz segmenta $[0, 1]$.

Problem proučavanja razdioba frekvencija za različite nizove statističkih podataka, na teorijskoj razini postaje problem proučavanja osobina f.r.v., tj. određenih realnih funkcija realne varijable koje posjeduju svojstva (2) i (3).

Odmah valja primjetiti da će svakom pojmu, koji je definiran pomoću razdiobe frekvencija, na teorijskoj razini odgovarati analogni pojam definiran pomoću vjerojatnosne razdiobe. Razdioba frekvencija proizlazi iz danoga konačnog niza statističkih podataka, dok se razdioba vjerojatnosti definira apstraktno-matematički bez pozivanja na empirijske činjenice, pa će i svi izvedeni pojmovi biti apstraktni.

Već je u I. poglavlju istaknuto da postoje određene razlike u tretiranju statističkih podataka za diskretno i za kontinuirano statističko obilježje. To se odražava i na izgradnju odgovarajućih matematičkih modela.

Diskretno statističko obilježje teorijski se interpretira kao ***diskretna slučajna varijabla*** X sa zadanim *skupom vrijednosti* $A = \{a_1, \ldots, a_r\}$ (može biti i $r = \infty$) i *pripadnim vjerojatnostima*

$$(4) \qquad p_j = \mathrm{P}(X = a_j) \geq 0, \quad j = 1, \ldots, r, \quad \sum_{j=1}^{r} p_j = 1\,.$$

Oznaka $\mathrm{P}(X = a_j)$, ili kraće $\mathrm{P}(a_j)$, označuje vjerojatnost da s.v. X poprimi vrijednost a_j. Broj p_j ima istu ulogu kao relativna frekvencija podataka a_j u razdiobi relativnih frekvencija opisanoj u I.2. Formulom (4) definirana je ***diskretna razdioba vjerojatnosti***.

Kao teorijski model za kontinuirano statističko obilježje uzima se ***kontinuirana slučajna varijabla*** X sa zadanom ***funkcijom gustoće vjerojatnosti*** (f.g.v.) $x \mapsto f(x) \geq 0$, $x \in \mathbf{R}$, čija su bitna svojstva

(5) $$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1,$$

(6) $$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x \geq 0, \quad a, b \in \mathbf{R}, \quad a < b.$$

Oznaka $\mathrm{P}(a \leq X \leq b)$ označuje vjerojatnost da s.v. X poprimi vrijednost koja nije manja od broja a i nije veća od broja b.

Iz (6) se razabire, stavljajući $b = a$, da za kontinuiranu s.v. vrijedi

(7) $$\mathrm{P}(X = a) = \mathrm{P}(a) = 0, \quad a \in \mathbf{R},$$

što znači da je vjerojatnost da kontinuirana s.v. X poprimi bilo koju pojedinačnu vrijednost a ($a \in \mathbf{R}$) jednaka nuli. Govori se da je zadanom f.g.v. $x \mapsto f(x)$ definirana ***kontinuirana razdioba vjerojatnosti***.

Kontinuirana s.v. matematički je model za one realne slučajne fenomene u kojima se kao rezultat mjerenja može dobiti bilo koji broj iz nekog intervala realnih brojeva, ili iz cijelog skupa $\mathbf{R}$. Statistička zakonitost održava se u različitoj gustoći rezultata mjerenja na pojedinim dijelovima (podskupovima) skupa $\mathbf{R}$ i upravo to se apstraktno-matematički izražava funkcijom gustoće vjerojatnosti.

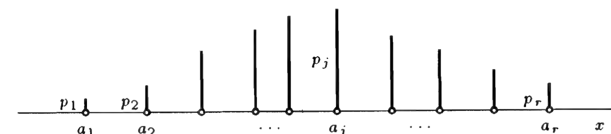
2. Diskretna razdioba vjerojatnosti

Funkcija razdiobe vjerojatnosti diskretne s.v. X može se, očigledno, zapisati u obliku

(8) $$F(x) = \sum_{\substack{j \\ a_j \leq x}} p_j, \quad x \in \mathbf{R},$$

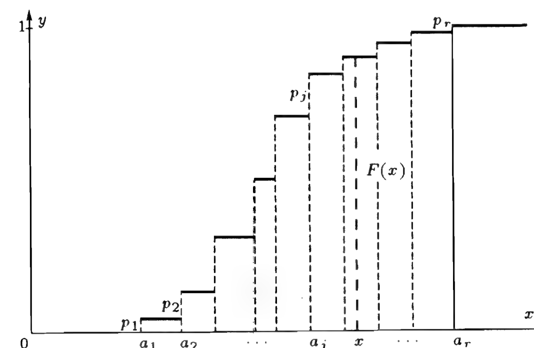
gdje je naznačeno da se zbraja po svim onim j-ovima za koje vrijedi $a_j \leq x$.

Geometrijski prikaz diskretne razdiobe vjerojatnosti može se načiniti tako da se na apscisnu os nanesu vrijednosti a_j ($j = 1, \ldots, r$), a kao pripadne ordinate uzmu odgovarajuće vjerojatnosti p_j (sl. 1).



Slika 1. Skica diskretne razdiobe vjerojatnosti

Drugi način grafičkog prikaza diskretne razdiobe vjerojatnosti jest taj da se skicira graf f.r.v. (sl. 2).



Slika 2. Skica grafa f.r.v. za diskretnu razdiobu

Sva važnija svojstva diskretne s.v. X mogu se izraziti pomoću njezinih vrijednosti a_j ($j = 1, \ldots, r$) i pripadnih vjerojatnosti $p_j = \mathrm{P}(X{=}a_j)$, tako da odgovarajuća f.r.v. (8) nema veliko praktično značenje.

Na temelju već istaknute analogije između razdiobe frekvencija i diskretne razdiobe vjerojatnosti mogu se na teorijskoj razini definirati pojmovi koji će biti analogni pojmovima definiranima u II. poglavlju za niz statističkih podataka. Tako se pojmu prosjeka niza statističkih podataka kao analogni teorijski pojam definira matematičko očekivanje diskretne slučajne varijable. Piše se

(9) $$\mathrm{E}[X] = \sum_{j=1}^{r} a_j p_j$$

i broj $\mathrm{E}[X]$ zove se ***matematičko očekivanje***, ili kraće ***očekivanje*** diskretne s.v. X. Govori se još da je $\mathrm{E}[X]$ *sredina* s.v. X. Umjesto $\mathrm{E}[X]$ često se, zbog kraćeg zapisa, upotrebljava oznaka μ, kada nije bitno da se istakne s.v. na koju se očekivanje odnosi.

Ako je $r = \infty$, onda se, dakako, postavlja i problem konvergencije reda u (9). Može se, naime, dogoditi i da $\mathrm{E}[X]$ ne postoji, ili da je beskonačno.

Jasno je da parametar $\mu = \mathrm{E}[X]$ karakterizira danu razdiobu vjerojatnosti u smislu lokacije, tj. njezin položaj na brojevnoj osi.

Teorijski analogon za pojam varijance niza statističkih podataka je pojam ***varijance*** ili ***disperzije*** diskretne s.v., koji se obično označuje sa $V[X]$, ili kraće σ^2, i definira formulom

$$(10) \qquad V[X] = \sigma^2 = \sum_{j=1}^{r}(a_j - \mu)^2 p_j .$$

Odmah se vidi da parametar σ^2 karakterizira danu diskretnu razdiobu vjerojatnosti u smislu raspršenja, odnosno rasipanja vrijednosti a_j diskretne s.v. X oko njezine sredine μ.

Iz (10) se lako dobiva formula

$$(11) \qquad V[X] = \sum_{j=1}^{r} a_j^2 p_j - \mu^2 = E[X^2] - (E[X])^2,$$

pri čemu je $E[X^2] = \sum_{j=1}^{r} a_j^2 p_j$. Formula (11) može se čitati i tako da se kaže da je varijanca jednaka očekivanju kvadrata minus kvadrat očekivanja s.v. X.

Sada je jasno da se na teorijskoj razini mogu definirati i analogoni statističkih momenata uvedenih u I.6. Tako se parametar

$$(12) \qquad \beta_k = \sum_{j=1}^{r} a_j^k p_j = E[X^k], \quad k = 0, 1, \ldots$$

zove *ishodišni* (*pomoćni*) *moment* k-tog reda, a parametar

$$(13) \qquad \mu_k = \sum_{j=1}^{r}(a_j - \mu)^k p_j, \quad k = 0, 1, \ldots$$

zove se ***centralni*** (*glavni*) ***moment*** k-tog reda diskretne s.v. X.

Definiraju se također i analogoni parametara oblika. Tako se parametar

$$(14) \qquad \kappa = \frac{\mu_3}{\sigma^3}$$

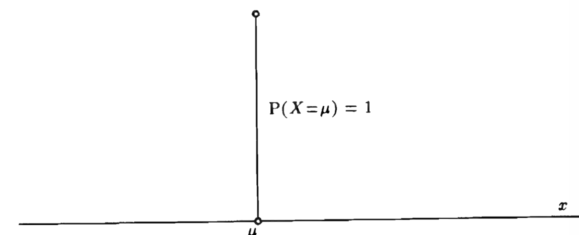
zove *koeficijent asimetrije*, a parametar

$$(15) \qquad \varepsilon = \frac{\mu_4}{\sigma^4} - 3$$

koeficijent spljoštenosti (*eksces*) dane diskretne razdiobe vjerojatnosti.

Svojstva i značenje parametara (očekivanje, varijanca, momenti i sl.) teorijske razdiobe vjerojatnosti slična su onima u empirijskim razdiobama relativnih frekvencija.

Ako je $V[X] = 0$, tada se govori o *degeneriranoj vjerojatnosnoj razdiobi*. Tada, naime, s.v. X poprima očekivanu vrijednost μ s vjerojatnošću jedan, tako da i nije riječ o "pravoj" s.v., već o konstanti μ.



Slika 3. Skica degenerirane razdiobe vjerojatnosti

3. Primjeri diskretnih razdioba vjerojatnosti

Smisao teorijskih vjerojatnosnih razdioba sastoji se u tome da pojedini tipovi teorijskih razdioba mogu poslužiti kao matematički model za mnoštvo različitih statističkih fenomena. Posebno su prikladni oni tipovi teorijskih razdioba koji se mogu jednostavno matematički opisati, tj. sama razdioba vjerojatnosti i najvažniji parametri mogu se izraziti jednostavnim formulama.

Kaže se da s.v. X ima ***binomnu razdiobu*** s parametrima m i p ($m \in \mathbf{N}, 0 < p < 1$) i piše se $X \sim B(m, p)$, ako je njezin skup vrijednosti $A = \{0, 1, \ldots, m\}$, a pripadne vjerojatnosti izražavaju se formulom

$$(16) \qquad p_j = P(X = j) = \binom{m}{j} p^j (1-p)^{m-j}, \quad j \in A.$$

Naziv binomna razdioba proizlazi iz činjenice da se u formuli (16) pojavljuje tzv. *binomni koeficijent* $\binom{m}{j}$ i da poznata formula *binomnog poučka* daje

$$(p+q)^m = \sum_{j=1}^{m} \binom{m}{j} p^j q^{m-j} = \sum_{j=1}^{m} p_j = 1, \quad q = 1 - p.$$

Lako se pokazuje da za najvažnije parametre binomne razdiobe vrijede formule

$$(17) \qquad E[X] = mp,$$

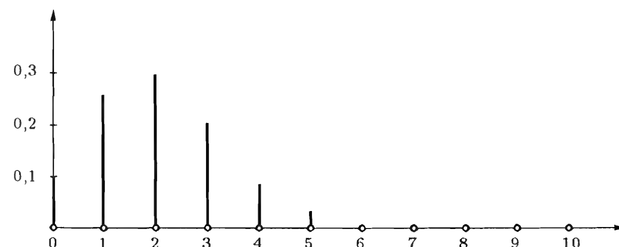
$$(18) \qquad V[X] = mp(1-p),$$

$$(19) \qquad \kappa = \frac{1-2p}{\sqrt{mp(1-p)}},$$

$$(20) \qquad \varepsilon = \frac{1-6p(1-p)}{mp(1-p)}.$$

Najtipičnija praktična situacija u kojoj se pojavljuje binomna razdioba jest m-struko ponavljanje slučajnog eksperimenta u kojem se uočava određeni događaj vjerojatnosti p i pritom promatra broj X nastupa (uspjeha) toga događaja (*Bernoullijeva shema*). Veličina X diskretna je s.v. binomne razdiobe $B(m,p)$.

Za $m = 1$ imamo tzv. *Bernoullijevu razdiobu* $B(1,p)$. Tada s.v. X ima skup vrijednosti $A = \{0,1\}$ i $X = 0$ označuje "neuspjeh" (nenastupanje događaja), a $X = 1$ "uspjeh" (nastupanje uočenog događaja) pri izvođenju danoga slučajnog eksperimenta.



Slika 4. Grafički prikaz B(10 ; 0,2)

Ako $X \sim B(10;0{,}2)$, onda je $P(X=0) = p_0 = \binom{10}{0} \cdot 0{,}2^0 \cdot (1-0{,}2)^{10} = 0{,}8^{10} \approx$ $\approx 0{,}11$, tj. vjerojatnost da se u deset ponavljanja slučajnog eksperimenta ne dobije nijedan uspjeh iznosi oko 11 %. Očekivani je broj uspjeha $E[X] = 10 \cdot 0{,}2 = 2$, uz varijancu $V[X] = 10 \cdot 0{,}2 \cdot 0{,}8 = 1{,}6$, odnosno standardnu devijaciju $\sigma = \sqrt{1{,}6} \approx 1{,}26$. Ova razdioba je pozitivno asimetrična ($\kappa \approx 0{,}47$).

Iz (19) se, inače, vidi da će binomna razdioba biti simetrična za $p = 0{,}5$, da će za $p < 0{,}5$ biti pozitivno asimetrična, a za $p > 0{,}5$ bit će negativno asimetrična.

Najveća vjerojatnost u $B(m,p)$ pripada vrijednosti $j_0 \in A$, za koju vrijedi

(21) $$p(m+1) - 1 \le j_0 \le p(m+1).$$

Za binomnu razdiobu sa sl. 4 vrijedi $p(m+1) = 0{,}2 \cdot 11 = 2{,}2$, tako da je $j_0 = 2$, a to je očigledno i iz sl. 4.

Za velike m i male p binomna razdioba $B(m,p)$ aproksimira se tzv. ***Poissonovom razdiobom***, pa se još govori da je to *razdioba "rijetkih događaja"*. Dokazuje se, naime, da vrijedi

$$\lim_{\substack{m\to\infty \\ p\to 0 \\ mp=\lambda}} \binom{m}{j} p^j (1-p)^{m-j} = \frac{\lambda^j}{j!}\exp(-\lambda), \quad \lambda > 0.$$

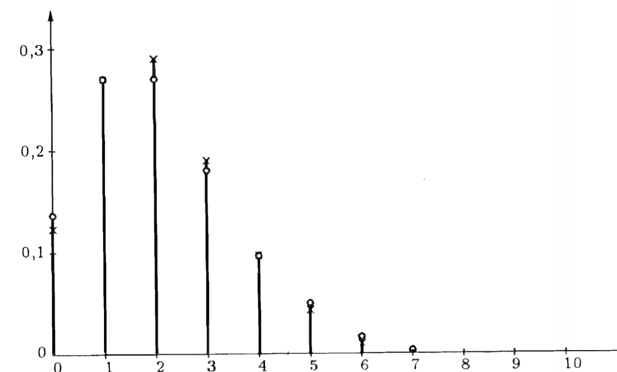
Budući da se, na temelju poznate činjenice da je $\exp(\lambda) = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$ ($\lambda \in \mathbf{R}$), lako dokazuje da je $\sum_{j=0}^{\infty} \frac{\lambda^j}{j!}\exp(-\lambda) = 1$, moguće je izreći sljedeću definiciju.

Kaže se da s.v. X ima Poissonovu razdiobu parametra λ i piše $X \sim Po(\lambda)$ ako je njezin skup vrijednosti $A = \{0,1,2,\dots\}$, a pripadne vjerojatnosti dane su formulom

(22) $$p_j = P(X=j) = \frac{\lambda^j}{j!}\exp(-\lambda), \quad j \in A.$$

Tablica 1.

j	Vjerojatnosti p_j u Po(2)	Vjerojatnosti p_j u B(20 ; 0,1)
0	0,135	0,122
1	0,271	0,270
2	0,271	0,285
3	0,181	0,190
4	0,090	0,090
5	0,036	0,032
6	0,012	0,009
7	0,003	0,002
8	0,001	0,000
9	0,000	0,000



Slika 5. Grafički prikaz Po(2) (kružići) i B(20 ; 0,1) (križići)

Za najvažnije parametre Poissonove razdiobe $Po(\lambda)$ vrijede formule

(23) $$E[X] = V[X] = \lambda,$$

(24) $$\kappa = \frac{1}{\sqrt{\lambda}}, \quad \varepsilon = \frac{1}{\lambda}.$$

Iz (24) se vidi da je Poissonova razdioba pozitivno asimetrična, a iz (22) se lako izvodi da najveća vjerojatnost pripada vrijednosti $j_0 \in A$, za koju vrijedi

(25) $$\lambda - 1 \le j_0 \le \lambda.$$

Ako je λ cijeli broj, onda, dakako, postoje dvije vrijednosti u skupu A kojima pripada maksimalna vjerojatnost.

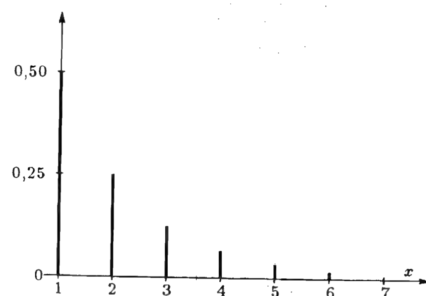
U tabl. 1. i na sl. 5. usporedno su prikazane Poissonova razdioba Po(2) i B(20 ; 0,1) ($mp = 2$), što zorno pokazuje odnos tih dviju vjerojatnosnih razdioba i smisao aproksimacije binomne razdiobe Poissonovom razdiobom.

Ako se pri ponavljanju nekoga slučajnog eksperimenta uoči određeni događaj vjerojatnosti p $(0 < p < 1)$ i zatim promatra veličina X koja označuje broj ponavljanja do prvog nastupa (uspjeha) tog događaja, onda je X diskretna slučajna varijabla sa skupom vrijednosti $A = \{1, 2, \ldots\}$ i pripadnim vjerojatnostima

(26) $$p_j = pq^{j-1}, \quad j \in A \quad (q = 1 - p).$$

Kaže se da diskretna s.v. X ima ***geometrijsku razdiobu*** parametra p. U dokazu jednakosti $\sum_{j=1}^{\infty} p_j = \sum_{j=1}^{\infty} pq^{j-1} = 1$ koristi se poznata činjenica da je $\sum_{j=1}^{\infty} q^{j-1}$ konvergentan geometrijski red, čija suma iznosi $\frac{1}{1-q} = \frac{1}{p}$. To omogućuje da se izvedu i formule za najvažnije parametre geometrijske razdiobe

(27) $$\mathrm{E}[X] = \frac{1}{p}, \quad \mathrm{V}[X] = \frac{q}{p^2}.$$

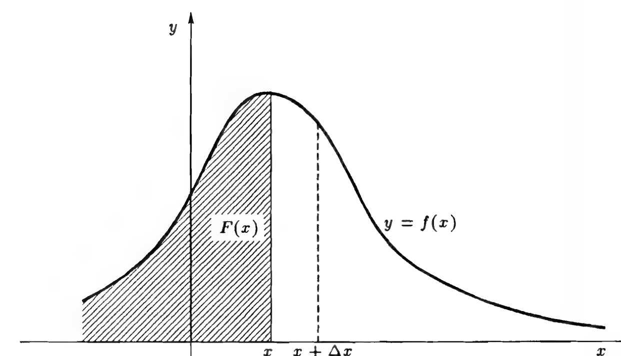


Slika 6. Grafički prikaz geometrijske razdiobe parametra $p = 0{,}5$

4. Kontinuirana razdioba vjerojatnosti

Za razliku od diskretne razdiobe, koja neposredno odražava svojstva empirijske razdiobe relativnih frekvencija nekog niza statističkih podataka, kontinuirana razdioba vjerojatnosti ne može se neposredno realizirati ni u kakvom nizu stvarnih statističkih podataka.

Kontinuirana razdioba vjerojatnosti najčešće se zorno prikazuje pomoću tzv. *krivulje razdiobe*, tj. grafa f.g.v.



Slika 7. Skica krivulje razdiobe

Prema (5) očigledno je da površina ispod krivulje razdiobe iznosi jedan, dok se iz (1) razabire da se f.r.v. može zapisati u obliku

(28) $$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t,$$

što znači da broj $F(x)$ izražava površinu ispod krivulje razdiobe, a iznad intervala $\langle -\infty, x]$.

Sada se vidi da se jednadžba (6) može zapisati i u obliku

(29) $$\mathrm{P}(a \le X \le b) = F(b) - F(a), \quad a, b \in \mathbf{R}, \quad a < b.$$

Deriviranjem jednadžbe (28) po x i vodeći računa o (29), dobiva se

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\mathrm{P}(x \le X \le x + \Delta x)}{\Delta x},$$

što za $f(x)$ opravdava naziv "gustoća vjerojatnosti" u točki x.

Iz (6) se, također, razabire da se vjerojatnost "padanja" rezultata mjerenja s.v. X u segment $[a, b]$ može zorno vidjeti kao površina ispod krivulje razdiobe, a iznad segmenta $[a, b]$. Iz ovoga se, nadalje, razabire da se za male Δx može pisati

$$\mathrm{P}(x \le X \le x + \Delta x) \approx f(x)\Delta x,$$

što označuje da je, teorijski gledano, udio onih rezultata mjerenja koji padaju u segment $[x, x + \Delta x]$ približno jednak $f(x)\Delta x$.

Definicija parametara kontinuirane razdiobe vjerojatnosti, kao što su matematičko očekivanje, varijanca, momenti itd., mora uzimati u obzir kontinuirani karakter promatrane s.v. X, tako da će odgovarajuće definicijske formule, umjesto suma, sadržavati integrale.

Matematičko očekivanje kontinuirane s.v. X definira se formulom

(30) $$\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}\,x,$$

dok se varijanca definira formulom

(31) $$\sigma^2 = \mathrm{V}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}\,x.$$

Ishodišni (pomoćni) momenti definiraju se formulom

(32) $$\beta_k = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}\,x,$$

a centralni (glavni) momenti formulom

(33) $$\mu_k = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,\mathrm{d}\,x.$$

Važno je primijetiti da se u formulama (30)-(33) pojavljuju nepravi integrali, kod kojih se postavlja pitanje konvergencije, tako da se može dogoditi da navedeni parametri za neke razdiobe vjerojatnosti i ne egzistiraju.

Stavi li se $\mathrm{E}[X^k] = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}\,x$, $k = 0, 1, \ldots$, odmah se vidi da se formule (30)-(33) mogu zapisati pomoću oznake E, pa je

(34) $$\mathrm{V}[X] = \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mu^2 = \mathrm{E}[X^2] - (\mathrm{E}[X])^2,$$

(35) $$\beta_k = \mathrm{E}[X^k], \quad \mu_k = \mathrm{E}[(X-\mu)^k], \quad k = 0, 1, \ldots$$

To omogućuje da se i za kontinuirane razdiobe definira koeficijent asimetrije i koeficijent spljoštenosti formulama (14) i (15).

Značenje očekivanja μ i varijance σ^2, kao najvažnijih parametara neke teorijske vjerojatnosne razdiobe, može se uočiti uz pomoć tzv. ***Čebiševljeve nejednakosti.*** Ako je, naime, c proizvoljan realan broj i za s.v. X (diskretna ili kontinuirana) vrijedi da je $\mathrm{E}[(X-c)^2] < \infty$, onda za svaki realni broj $\delta > 0$ vrijedi

(36) $$\mathrm{P}(|X-c| \geq \delta) \leq \frac{1}{\delta^2}\mathrm{E}[(X-c)^2].$$

Uzme li se $c = \mu$, tada je $\mathrm{E}[(X-c)^2] = \mathrm{E}[(X-\mu)^2] = \sigma^2$ i (36) postaje

$$\mathrm{P}(|X-\mu| \geq \delta) \leq \frac{\sigma^2}{\delta^2}.$$

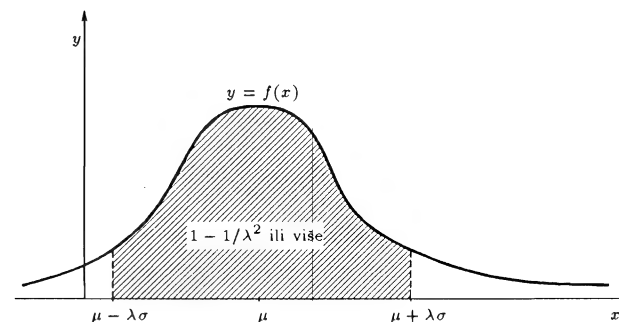
Stavi li se još $\delta = \lambda\sigma$ ($\lambda > 0$), dobiva se

$$\mathrm{P}(|X-\mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2},$$

što se može pisati i kao

(37) $$\mathrm{P}(\mu - \lambda\sigma < X < \mu + \lambda\sigma) \geq 1 - \frac{1}{\lambda^2}.$$

Relacija (37) pokazuje, kao što se vidi i na sl. 8, da je u svakoj vjerojatnosnoj razdiobi interval $\langle \mu - \lambda\sigma, \mu + \lambda\sigma \rangle$ "opterećen" bar vjerojatnošću $1 - \frac{1}{\lambda^2}$, odnosno da izvan toga intervala ne ostaje više od $\frac{1}{\lambda^2}$ vjerojatnosnog opterećenja.



Slika 8. Grafička interpretacija Čebiševljeve nejednakosti

Posebno, uzme li se $\lambda = 3$, dobiva se da je interval $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ opterećen bar vjerojatnošću $1 - \frac{1}{9} = \frac{8}{9}$, odnosno da u svakom teorijskom modelu udio onih vrijednosti s.v. X koja padaju u interval $\langle \mu - 3\sigma, \mu + 3\sigma \rangle$ iznosi bar $\frac{8}{9} \approx 90\,\%$.

Čebiševljeva nejednakost teorijski je analogon relacije (16) iz II.4, koja se odnosi na empirijsku razdiobu frekvencija.

U II.2. uveden je pojam medijana, a u II.5. pojam kvartila za niz statističkih podataka, pa se prirodno nameće ideja da se i za teorijsku vjerojatnosnu razdiobu definiraju analogni pojmovi. Grubo rečeno, medijan vjerojatnosne razdiobe bit će ona vrijednost (točka) koja razdiobu vjerojatnosti dijeli na dva jednaka dijela od po 0,5 (ukupno raspodijeljena vjerojatnost iznosi 1). Uzme li se u obzir značenje f.r.v. definirane formulom (1), odmah se vidi da je ***medijan*** M dane vjerojatnosne razdiobe ona vrijednost ($M \in \mathbf{R}$) za koju vrijedi

(38) $$F(M) = 0{,}5.$$

Slično se definira i općenitiji pojam ***kvantila*** x_p ***reda*** p ($0 < p < 1$) vjerojatnosne razdiobe zadane funkcijom razdiobe vjerojatnosti F, kao ona vrijednost ($x_p \in \mathbf{R}$) za koju vrijedi

(39) $$F(x_p) = p.$$

Očigledno je $M = x_{0,5}$, tj. medijan je kvantil reda 0,5.

Kvantili $x_{0,25}$ i $x_{0,75}$ zovu se *kvartili*, a veličina

(40) $$\delta_2 = x_{0,75} - x_{0,25}$$

zove se *interkvartilni raspon* zadane vjerojatnosne razdiobe.

Ako je riječ o *simetričnoj razdiobi* vjerojatnosti sa središtem simetrije u točki μ ($\mu \in \mathbf{R}$), tj. ako pripadna f.g.v. zadovoljava uvjet $f(\mu - x) = f(\mu + x)$, za svaki $x \in \mathbf{R}$, onda se medijan M poklapa s očekivanjem $\mathrm{E}[X] = \mu$ (ako postoji) i za kvantile općenito vrijedi

(41) $$x_p = 2\mu - x_{1-p}.$$

Analogon svojstva medijana niza statističkih podataka iskazanog relacijom (10) iz II.2. sada glasi

(42) $$\min_{c \in \mathbf{R}} \mathrm{E}[|X - c|] = \mathrm{E}[|X - M|].$$

Formulom (42) izriče se činjenica da se minimalno očekivanje apsolutne razlike (udaljenosti) između s.v. X i realnog broja c postiže onda kada se uzme $c = M$.

5. Primjeri kontinuiranih razdioba vjerojatnosti

Konkretna teorijska kontinuirana razdioba vjerojatnosti obično se zadaje svojom funkcijom gustoće vjerojatnosti. Najvažniji model teorijske razdiobe vjerojatnosti uopće je ***normalna*** ili ***Gaussova razdioba***.

Za kontinuiranu s.v. X kaže se da ima normalnu razdiobu s parametrima μ i σ^2 i piše $X \sim \mathrm{N}(\mu, \sigma^2)$, ako je njezina f.g.v. zadana formulom

(43) $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Može se dokazati da je

(44) $$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \mathrm{d}x = 1,$$

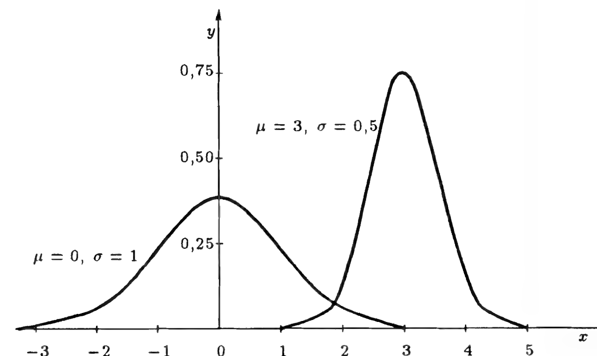
a također i formule

(45) $$\mathrm{E}[X] = \mu, \qquad \mathrm{V}[X] = \sigma^2, \qquad \kappa = 0, \qquad \varepsilon = 0.$$

Odmah se vidi da parametri μ i σ^2 normalne razdiobe imaju važno statističko značenje očekivanja (sredine) i varijance (disperzije) s.v. X.

Krivulja normalne razdiobe simetrična je u odnosu na pravac $x = \mu$, ima maksimum $\frac{1}{\sigma\sqrt{2\pi}}$ za $x = \mu$, dok za $x = \mu - \sigma$ i $x = \mu + \sigma$ ima točke infleksije.

Može se reći da je normalna razdioba $\mathrm{N}(\mu, \sigma^2)$ teorijski model za simetričnu razdiobu rezultata mjerenja, čija gustoća zvonoliko opada s udaljavanjem od središta simetrije μ (sl. 9).



Slika 9. Primjeri krivulja normalne razdiobe

Iz (28) proizlazi da se f.r.v. za $\mathrm{N}(\mu, \sigma^2)$ izražava formulom

(46) $$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] \mathrm{d}t.$$

Integral na desnoj strani u (46) nije elementarno rješiv i zato se u računima s normalnom razdiobom upotrebljavaju tablice, gdje su navedene vrijednosti f.r.v. *standardne* ili *jedinične normalne razdiobe* $\mathrm{N}(0,1)$, za koju je $\mu = 0$ i $\sigma^2 = 1$. Pripadna f.g.v. obično se označuje sa φ, a f.r.v. sa Φ, tako da je

(47) $$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t.$$

Funkcija φ je parna, tj. vrijedi $\varphi(-x) = \varphi(x)$, dok za funkciju Φ vrijedi

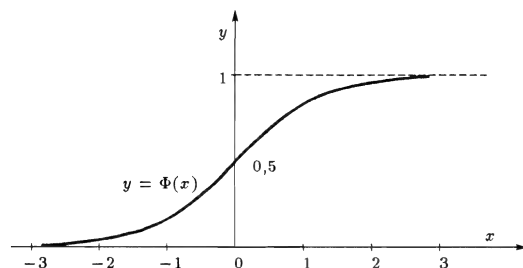
(48) $$\Phi(-x) = 1 - \Phi(x), \qquad \Phi(0) = 0{,}5.$$

Stoga je dovoljno tablično prikazati vrijednosti tih funkcija za $x \geq 0$.

Veza između $\mathrm{N}(\mu, \sigma^2)$ i $\mathrm{N}(0,1)$ uspostavlja se formulama

(49) $$f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right), \qquad F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

tako da za s.v. $X \sim \mathrm{N}(\mu, \sigma^2)$ i $a < b$ vrijedi



Slika 10. Graf funkcije Φ

$$\text{(50)}\qquad \mathrm{P}(a \le X \le b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Posebno, za proizvoljno $\lambda > 0$, iz (50) se dobiva

$$\text{(51)}\qquad \mathrm{P}(\mu - \lambda\sigma \le X \le \mu + \lambda\sigma) = 2\Phi(\lambda) - 1.$$

Uzme li se $\lambda = 3$, iz (51) i tablice za funkciju Φ (v. tabl. III. u Dodatku), proizlazi

$$\text{(52)}\qquad \mathrm{P}(\mu - 3\sigma \le X \le \mu + 3\sigma) \approx 0{,}9973,$$

iz čega se vidi da, iako teorijski gledano normalna razdioba svakome realnom broju pridružuje pozitivnu gustoću vjerojatnosti, praktički gledano interval $[\mu-3\sigma, \mu+3\sigma]$ obuhvaća gotovo sveukupno (99,73 %) vjerojatnosno opterećenje koje iznosi jedan.

Druga važna dvoparametarska familija teorijskih kontinuiranih razdioba vjerojatnosti, koja se rabi kao matematički model za one realne statističke fenomene u kojima se kao rezultati mjerenja mogu dobiti samo nenegativni brojevi, je ***gama-razdioba***. Kaže se da s.v. X ima gama-razdiobu s parametrima α i β $(\alpha > 0, \beta > 0)$ i piše $X \sim \mathrm{G}(\alpha, \beta)$, ako pripadna f.g.v. glasi

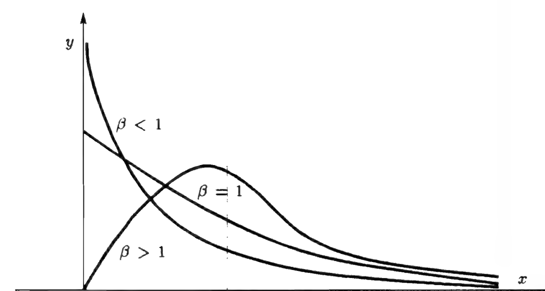
$$\text{(53)}\qquad f(x) = \begin{cases} 0, & \text{za } x \le 0 \\ \dfrac{\alpha^\beta}{\Gamma(\beta)}\, x^{\beta-1}\exp(-\alpha x), & \text{za } x > 0, \end{cases}$$

gdje je $\Gamma(\beta) = \int_0^\infty t^{\beta-1}\exp(-t)\,\mathrm{d}t$, tj. $\beta \mapsto \Gamma(\beta)$, $\beta > 0$, jest tzv. *gama-funkcija*, čije se vrijednosti obično tablično prikazuju (v. tabl. IV. u Dodatku).

Za $\beta = 1$ dobiva se ***eksponencijalna razdioba*** parametra α i tada se piše $X \sim \mathrm{Ex}(\alpha)$. Iz (53) se razabire da pripadna f.g.v. glasi

$$\text{(54)}\qquad f(x) = \begin{cases} 0, & \text{za } x \le 0 \\ \alpha\exp(-\alpha x), & \text{za } x > 0. \end{cases}$$

Eksponencijalna razdioba najčešće se pojavljuje kao matematički model za opisivanje slučajnog vijeka trajanja određenoga elektrotehničkog ili nekoga drugog proizvoda (žarulja, otpornik, kondenzator i sl.).



Slika 11. Skica različitih krivulja gama-razdiobe

Ako se u gama-razdiobi specificiraju parametri α i β tako da se stavi $\alpha = 0{,}5$ i $\beta = \frac{n}{2}$ $(n \in \mathbf{N})$, onda se govori o ***hikvadrat-razdiobi sa n stupnjeva slobode*** i piše $X \sim \chi^2(n)$.

Da bi se u primjenama izbjegla složena računanja u vezi s gama-razdiobama, izrađene su tablice gdje su navedene vrijednosti za f.r.v. uz određene vrijednosti parametara.

Odgovarajućim izvodima može se dokazati da za s.v. $X \sim \mathrm{G}(\alpha, \beta)$ vrijedi

$$\text{(55)}\qquad \mathrm{E}[X] = \frac{\beta}{\alpha},\quad \mathrm{V}[X] = \frac{\beta}{\alpha^2},\quad \kappa = \frac{2}{\sqrt{\beta}},\quad \varepsilon = \frac{6}{\beta}.$$

Iz ovoga odmah proizlazi da za $X \sim \mathrm{Ex}(\alpha)$ vrijedi

$$\text{(56)}\qquad \mathrm{E}[X] = \frac{1}{\alpha},\quad \mathrm{V}[X] = \frac{1}{\alpha^2},\quad \kappa = 2,\quad \varepsilon = 6,$$

dok za $X \sim \chi^2(n)$ vrijedi

$$\text{(57)}\qquad \mathrm{E}[X] = n,\quad \mathrm{V}[X] = 2n,\quad \kappa = \sqrt{\frac{8}{n}},\quad \varepsilon = \frac{12}{n}.$$

Ako se kao moguće vrijednosti mjerenja u nekom slučajnom eksperimentu pojavljuju samo brojevi iz intervala $\langle 0, 1\rangle$, onda se, kao odgovarajući matematički model, obično uzima neka od teorijskih razdioba iz dvoparametarske familije vjerojatnosnih razdioba koja se zove ***beta-razdioba***. Kaže se da s.v. X ima beta-razdiobu s parametrima α i β $(\alpha > 0, \beta > 0)$ ako je njezina f.g.v. zadana formulom

$$\text{(58)}\qquad f(x) = \begin{cases} 0, & \text{za } x \le 0 \text{ i } x \ge 1 \\ \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, & \text{za } 0 < x < 1. \end{cases}$$

Pokazuje se da je

$$\text{(59)}\qquad \mathrm{E}[X] = \frac{\alpha}{\alpha+\beta},\qquad \mathrm{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

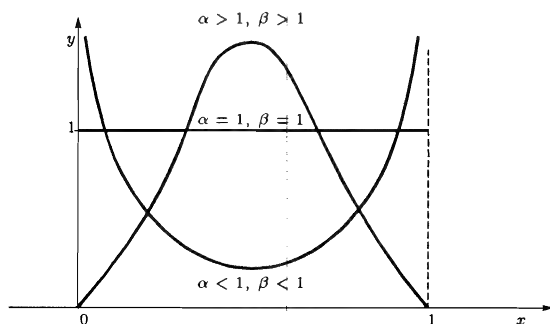
Za $\alpha = 1$ i $\beta = 1$, iz (58) proizlazi

(60) $$f(x) = \begin{cases} 0, & \text{za } x \leq 0 \text{ i } x \geq 1 \\ 1, & \text{za } 0 < x < 1 \end{cases}$$

i tada se govori o ***jednolikoj (uniformnoj) razdiobi nad intervalom*** $\langle 0,1\rangle$ i piše $X \sim \mathrm{U}(0,1)$. Uniformna razdioba $\mathrm{U}(0,1)$ teorijski opisuje razdiobu rezultata slučajnih mjerenja u kojima se svaki broj iz intervala $\langle 0,1\rangle$ pojavljuje s istom šansom. Često se, umjesto o jednolikoj razdiobi nad intervalom $\langle 0,1\rangle$, govori o *jednolikoj razdiobi* U(0,1) *nad segmentom* $[0,1]$, između čega nema bitnih razlika.

Za $X \sim \mathrm{U}(0,1)$ vrijedi

(61) $$\mathrm{E}[X] = \frac{1}{2}, \quad \mathrm{V}[X] = \frac{1}{12}, \quad \kappa = 0, \quad \varepsilon = -1{,}2.$$



Slika 12. Skica različitih krivulja beta-razdiobe

6. Funkcije slučajne varijable

Mnoge praktične situacije zahtijevaju da se, umjesto s izmjerenima statističkim podacima, radi s transformiranim podacima pomoću određene realne funkcije realne varijable. Ako je originalni niz podataka $x_1, \ldots, x_n$, niz transformiranih podataka bit će $y_1, \ldots, y_n$, gdje je $y_i = h(x_i)$ $(i = 1, \ldots, n)$, pri čemu je h zadana funkcija (transformacija). Odmah se postavlja pitanje određivanja ovisnosti (formula) koje povezuje razdiobu frekvencija i odgovarajuće parametre izvornih podataka i transformiranih podataka.

Na teorijskoj razini taj se problem svodi na određivanje veza između zadane slučajne varijable X, njezine vjerojatnosne razdiobe i pripadnih parametara, s jedne strane, i s.v. $Y = h(X)$, odnosno njezine razdiobe vjerojatnosti i pripadnih parametara, s druge strane.

Ako je, na primjer, riječ o tzv. *afinoj transformaciji* zadanoj formulom $h(x) = ax + b$ $(a \neq 0)$, i ako s.v. X pripada f.r.v. F, onda slučajnoj varijabli $Y = aX + b$ pripada funkcija razdiobe vjerojatnosti

(62) $$\mathrm{G}(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(aX + b \leq y) = \mathrm{P}\left(X \leq \frac{y-b}{a}\right) = F\left(\frac{y-b}{a}\right).$$

Pokazuje se, također, da vrijedi

(63) $$\mathrm{E}[Y] = \mathrm{E}[aX + b] = a\mathrm{E}[X] + b\,,$$

(64) $$\mathrm{V}[Y] = \mathrm{V}[aX + b] = a^2\mathrm{V}[X]\,.$$

Ako je X kontinuirana s.v. s pripadnom f.g.v. f, onda je $Y = aX + b$ također kontinuirana s.v. s pripadnom funkcijom gustoće vjerojatnosti

(65) $$g(y) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right).$$

Ako još $X \sim \mathrm{N}(\mu, \sigma^2)$, onda $Y = aY + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$, tj. afinom transformacijom normalne razdiobe dobiva se opet normalna razdioba.

Ako $X \sim \mathrm{U}(0,1)$, onda je $Y = aX + b$ kontinuirana s.v. kojoj pripada ***jednolika (uniformna) razdioba nad intervalom*** $\langle a_0, b_0\rangle$, gdje je $a_0 = b$ i $b_0 = a + b$, za $a > 0$, a za $a < 0$ je $a_0 = a + b$ i $b_0 = b$. Piše se $Y \sim \mathrm{U}(a_0, b_0)$ i tada je

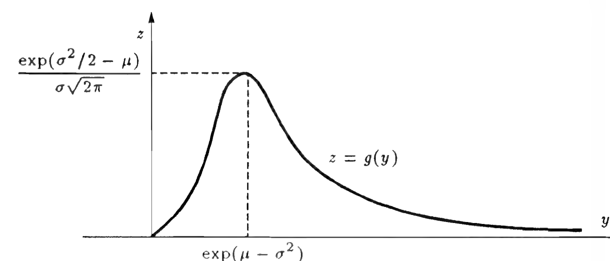
(66) $$g(y) = \begin{cases} 0, & \text{za } y \leq a_0 \text{ i } y \geq b_0 \\ \dfrac{1}{b_0 - a_0}, & \text{za } a_0 < y < b_0, \end{cases},$$

(67) $$\mathrm{E}[Y] = \frac{1}{2}(a_0 + b_0), \quad \mathrm{V}[Y] = \frac{1}{12}(b_0 - a_0)^2\,.$$

Pretpostavi li se, pak, da $X \sim \mathrm{N}(\mu, \sigma^2)$ i da je $Y = \exp(X)$, tj. da se X podvrgava eksponencijalnoj transformaciji, pokazuje se da je Y kontinuirana s.v. s pripadnom funkcijom gustoće vjerojatnosti

(68) $$g(y) = \begin{cases} 0, & \text{za } y \leq 0 \\ \dfrac{1}{y\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}\left(\dfrac{\ln y - \mu}{\sigma}\right)^2\right], & \text{za } y > 0. \end{cases}$$

Vjerojatnosna razdioba koju karakterizira (68) zove se ***lognormalna razdioba*** s parametrima μ i σ^2. Piše se $Y \sim \mathrm{LN}(\mu, \sigma^2)$.



Slika 13. Skica krivulje lognormalne razdiobe

Osnovni parametri lognormalne razdiobe izraženi su formulama

(69) $$\mathrm{E}[Y] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad \mathrm{V}[Y] = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1].$$

Očigledno je da se lognormalna razdioba, slično kao i gama-razdioba, može uzeti kao teorijski model za one praktične statističke fenomene gdje se kao rezultati mjerenja pojavljuju samo nenegativni realni brojevi.

Ako se u vezi sa s.v. X promotri s.v. $Y = (X - c)^2$, gdje je c proizvoljan realan broj, onda se vidi da je

$$\mathrm{E}[Y] = \mathrm{E}[(X-c)^2] = \mathrm{E}[((X-\mu)+(\mu-c))^2] = \mathrm{E}[(X-\mu)^2] + (\mu - c)^2,$$

a iz toga odmah proizlazi da je

(70) $$\min_{c\in\mathbf{R}} \mathrm{E}[(X-c)^2] = \mathrm{E}[(X-\mu)^2] = \mathrm{V}[X].$$

Time je iskazano da je očekivani kvadrat udaljenosti s.v. X od nekoga realnog broja c minimalan onda ako je $c = \mu = \mathrm{E}[X]$ i jednak je upravo varijanci $\mathrm{V}[X] = \sigma^2$. Relacija (70) teorijski je analogon relacije (4) iz II.1.

Zadaci

1. Ako je $A = \{a_1, \ldots, a_r\}$ i $p_j = \frac{1}{r}$ $(j = 1, \ldots, r)$, onda se govori o *jednolikoj (uniformnoj) razdiobi na skupu A*.

a) Napišite formulu za pripadnu f.r.v. i skicirajte njezin graf ako je $a_j = j$.
b) Izvedite formule za $\mathrm{E}[X]$, $\mathrm{V}[X]$, κ i ε.

2. Napišite formulu za f.r.v. jednolike razdiobe nad intervalom $\langle a, b\rangle$ $(a < b)$ i izvedite odgovarajuće formule za očekivanje, varijancu, koeficijent asimetrije i koeficijent spljoštenosti.

3. Dokažite da je:

a) $\mathrm{E}[X - \mu] = 0$, $\mu = \mathrm{E}[X]$,
b) $\mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$.

4. Izvedite formulu

$$\mu_k = \sum_{i=1}^{k} (-1)^{k-i}\binom{k}{i}\beta_k \mu^{k-i}, \quad k = 0, 1, 2, \ldots,$$

gdje je μ_k centralni, a β_k ishodišni moment k-tog reda.

5. Dokažite da za binomnu razdiobu $B(r, p)$ vrijedi rekurzivna formula

$$p_j = \frac{r - j + 1}{j} \cdot \frac{p}{1-p} \cdot p_{j-1}, \quad j = 1, 2, \ldots, r.$$

6. Izvedite formule (17)–(21).

7. Dokažite da za Poissonovu razdiobu $\mathrm{Po}(\lambda)$ vrijedi rekurzivna formula

$$p_j = \frac{\lambda}{j} p_{j-1}, \quad j = 1, 2, \ldots.$$

8. Izvedite formule (23)–(25).

9. Izvedite formule (27).

10. Dokažite Čebiševljevu nejednakost za:

a) diskretnu s.v.,
b) kontinuiranu s.v.

11. Dokažite da za kontinuiranu s.v. X vrijedi

$$\min_{c\in\mathbf{R}} \mathrm{E}[|X - c|] = \mathrm{E}[|X - M|],$$

gdje je M pripadni medijan.

12. Izvedite formule za ishodišne i centralne momente:

a) normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$,
b) eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$.

13. Izvedite formule (48) oslanjajući se na formulu (44).

14. Dokažite valjanost formula (48)–(51).

15. Izvedite formule (55).

16. Izvedite formule (61).

17. Dokažite valjanost formula (63) i (64).

18. Izvedite relaciju (70).

19. Izvedite formulu za medijan:

a) normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$,
b) eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$,
c) uniformne razdiobe $\mathrm{U}(a, b)$.

V. Teorijska interpretacija višedimenzionalnih statističkih obilježja

1. Dvodimenzionalna razdioba vjerojatnosti

Kao što su u IV. poglavlju prikazani elementi matematičke teorije koja objašnjava različite fenomene u vezi s jednodimenzionalnim statističkim obilježjima, sada će se, analogno tome najprije prikazati elementi matematičke teorije koja objašnjava fenomene u vezi s dvodimenzionalnim statističkim obilježjima, a zatim će se navesti i određeni pojmovi i metode ***teorije višedimenzionalnih slučajnih varijabli***. Teorija višedimenzionalnih s.v., kao što je već uočeno kod jednodimenzionalnih s.v., temelji se na spoznaji da se, pri simultanom mjerenju dviju ili više različitih veličina (statističkih obilježja) i mnogostrukom ponavljanju tih mjerenja, dobivaju brojčani rezultati u kojima se uočavaju određene statističke zakonitosti. Matematički model za apstraktno teorijsko opisivanje takvih pojava zove se ***slučajni vektor*** (s.vk.), čije ***komponente*** su slučajne varijable.

Ako se simultano promatraju dva statistička obilježja, onda se govori o slučajnom vektoru (X, Y), čije su komponente s.v. X i Y.

Pri empirijskim razmatranjima u III. poglavlju govorilo se o dvodimenzionalnoj razdiobi relativnih frekvencija, dok će se u matematičkoj teoriji govoriti o ***dvodimenzionalnoj razdiobi vjerojatnosti***. Tako se za s.vk. (X, Y) pripadna f.r.v. definira formulom

(1) $$F(x,y) = \mathrm{P}(X \le x, Y \le y), \quad x, y \in \mathbf{R}.$$

To je, dakle, realna funkcija dviju realnih varijabli i $F(x,y)$ je realan broj koji označuje teorijsku vjerojatnost da se, pri simultanom mjerenju slučajnih varijabli X i Y, dobije za X vrijednost koja ne premašuje broj x i za Y vrijednost koja ne premašuje broj y.

Lako se uviđa da vrijedi

(2) $F(-\infty, y) = F(x, -\infty) = 0$,

(3) $F(\infty, \infty) = 1$,

(4) $x_1 < x_2 \Longrightarrow F(x_1, y) \le F(x_2, y), \quad y_1 < y_2 \Longrightarrow F(x, y_1) \le F(x, y_2)$,

(5) $\mathrm{P}(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1)$.

Funkcija $x \mapsto F_1(x) = F(x, \infty) = \mathrm{P}(X \le x)$, $x \in \mathbf{R}$, zove se ***marginalna f.r.v. komponente*** X, a funkcija $y \mapsto F_2(y) = F(\infty, y) = \mathrm{P}(Y \le y)$, $y \in \mathbf{R}$, zove se ***marginalna f.r.v. komponente*** Y.

Marginalne vjerojatnosne razdiobe teorijski opisuju statističko ponašanje svake slučajne varijable X i Y posebno.

Ako vrijedi $F(x,y) = F_1(x)F_2(y)$, za sve $x, y \in \mathbf{R}$, onda se kaže da su X i Y ***stohastički nezavisne*** s.v.

Radi jednostavnosti pisanja i govorenja, umjesto stohastički nezavisne obično se piše i govori samo ***nezavisne*** s.v.

Ako su X i Y nezavisne s.v., onda njihovo simultano proučavanje ne daje nikakve nove informacije s obzirom na zasebno proučavanje svake od njih.

2. Diskretna dvodimenzionalna razdioba vjerojatnosti

Neka su $A = \{a_1, \dots, a_r\}$ i $B = \{b_1, \dots, b_s\}$ zadani diskretni skupovi (može biti i $r = \infty$ i $s = \infty$) realnih brojeva i $p_{ij} \ge 0$ $\left(\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1\right)$ zadani brojevi. Broj p_{ij} interpretira se kao vjerojatnost da s.vk. (X, Y) poprimi vrijednost $(a_i, b_j) \in A \times B$. Piše se

(6) $$p_{ij} = \mathrm{P}(X = a_i, Y = b_j), \quad i = 1, \dots, r, \quad j = 1, \dots, s,$$

i govori da je (X, Y) diskretni s.vk. sa skupom vrijednosti $A \times B$ i pripadnim vjerojatnostima p_{ij}. Kaže se još da slučajnom vektoru (X, Y) pripada ***diskretna dvodimenzionalna vjerojatnosna razdioba*** zadana formulom (6). To je matematički model za diskretno dvodimenzionalno statističko obilježje razmotreno u III. poglavlju. Umjesto empirijskih relativnih frekvencija ovdje se pojavljuju apstraktne teorijske veličine – vjerojatnosti p_{ij}.

Uzme li se, na primjer, $A = B = \{0, 1, \dots, m\}$ $(m \in \mathbf{N})$ i stavi

(7) $$p_{ij} = \mathrm{P}(X = i, Y = j) = \begin{cases} 0, & \text{za } i + j > m \\ \dfrac{m!}{i!j!(m-i-j)!} p_1^i p_2^j (1 - p_1 - p_2)^{m-i-j}, & \text{za } i + j \le m, \end{cases}$$

gdje su p_1 i p_2 $(p_1 > 0,\ p_2 > 0,\ p_1 + p_2 < 1)$ zadani parametri, onda se kaže da diskretni s.vk. (X, Y) ima *trinomnu razdiobu* $\mathrm{B}(m, p_1, p_2)$ s parametrima m, p_1 i p_2 i piše $(X, Y) \sim \mathrm{B}(m, p_1, p_2)$. Trinomna razdioba u određenom je smislu generalizacija binomne razdiobe razmotrene u IV.3.

Marginalne razdiobe diskretnog s.vk. (X, Y) također su diskretne razdiobe. Tako komponenta X ima diskretnu vjerojatnosnu razdiobu sa skupom vrijednosti A i pripadnim vjerojatnostima

(8) $$p_i = \mathrm{P}(X = a_i) = \sum_{j=1}^{s} p_{ij}, \quad i = 1, \dots, r,$$

a komponenta Y ima diskretnu vjerojatnosnu razdiobu sa skupom vrijednosti B i pripadnim vjerojatnostima

$$q_j = \mathrm{P}(Y = b_j) = \sum_{i=1}^{r} p_{ij}, \quad j = 1, \ldots, s. \tag{9}$$

Uvjet nezavisnosti slučajnih varijabli X i Y sada glasi

$$p_{ij} = p_i\, q_j, \quad i = 1, \ldots, r, \quad j = 1, \ldots, s. \tag{10}$$

Budući da su marginalne razdiobe jednodimenzionalne vjerojatnosne razdiobe, mogu se definirati parametri

$$\mu_1 = \mathrm{E}[X] = \sum_{i=1}^{r} a_i p_i, \quad \sigma_1^2 = \mathrm{V}[X] = \sum_{i=1}^{r} (a_i - \mu_1)^2 p_i,$$

$$\mu_2 = \mathrm{E}[Y] = \sum_{j=1}^{s} b_j q_j, \quad \sigma_2^2 = \mathrm{V}[Y] = \sum_{j=1}^{s} (b_j - \mu_2)^2 q_j.$$

Može se pokazati (v. zad. 3) da iz $(X, Y) \sim \mathrm{B}(m, p_1, p_2)$ proizlazi da $X \sim \mathrm{B}(m, p_1)$ i $Y \sim \mathrm{B}(m, p_2)$, tj. da su marginalne razdiobe u trinomnoj razdiobi, binomne razdiobe. Iz toga slijedi da je $\mathrm{E}[X] = mp_1$, $\mathrm{E}[Y] = mp_2$, $\mathrm{V}[X] = mp_1(1 - p_1)$ i $\mathrm{V}[Y] = mp_2(1 - p_2)$.

Ako je u danoj diskretnoj dvodimenzionalnoj razdiobi vjerojatnosti $q_j > 0$, onda se može definirati jednodimenzionalna diskretna vjerojatnosna razdioba sa skupom vrijednosti A i pripadnim vjerojatnostima

$$p_{i/j} = \frac{p_{ij}}{q_j}, \quad i = 1, \ldots, r, \tag{11}$$

koja se zove ***uvjetna razdioba vjerojatnosti*** komponente X uz fiksiranu vrijednost b_j komponente Y. Veličina definirana formulom

$$\mathrm{E}[X/b_j] = \sum_{i=1}^{r} a_i p_{i/j} \tag{12}$$

zove se ***uvjetno očekivanje*** komponente X uz fiksiranu vrijednost b_j komponente Y.

Zamjenom uloga komponenti X i Y i polazeći od $p_i > 0$, može se definirati jednodimenzionalna diskretna vjerojatnosna razdioba sa skupom vrijednosti B i pripadnim vjerojatnostima

$$q_{j/i} = \frac{p_{ij}}{p_i}, \quad j = 1, \ldots, s, \tag{13}$$

koja se zove uvjetna razdioba komponente Y uz fiksiranu vrijednost a_j komponente X. Veličina definirana formulom

$$\mathrm{E}[Y/a_j] = \sum_{j=1}^{s} b_j q_{j/i} \tag{14}$$

zove se uvjetno očekivanje komponente Y uz fiksiranu vrijednost a_j komponente X.

Za trinomnu razdiobu $\mathrm{B}(m, p_1, p_2)$, na primjer, uvjetne razdiobe su binomne razdiobe. Tako je uvjetna razdioba komponente X uz fiksirano j ($j \in \{0, 1, \ldots, m\}$) binomna razdioba $\mathrm{B}\left(m - j, \dfrac{p_1}{1 - p_2}\right)$, iz čega proizlazi da je odgovarajuće uvjetno očekivanje $\mathrm{E}[X/j] = (m - j)\dfrac{p_1}{1 - p_2}$. Analogno je binomna razdioba $\mathrm{B}\left(m - i, \dfrac{p_2}{1 - p_1}\right)$ uvjetna razdioba komponente Y uz fiksiranu vrijednost i komponente X, pa je pripadno uvjetno očekivanje $\mathrm{E}[Y/i] = (m - i)\dfrac{p_2}{1 - p_1}$.

3. Kontinuirana dvodimenzionalna razdioba vjerojatnosti

Na temelju analogije s jednodimenzionalnom kontinuiranom vjerojatnosnom razdiobom (IV.1. i IV.4) izgrađuje se i matematički model za kontinuirano dvodimenzionalno statističko obilježje. Govori se o kontinuiranome slučajnom vektoru (X, Y) sa zadnom funkcijom gustoće vjerojatnosti $(x, y) \mapsto f(x, y)$, $(x, y) \in \mathbf{R}^2$, pri čemu vrijedi $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1$, kao o teorijskom pojmu kojim se tumače stvarni fenomeni pri simultanom mjerenju dvaju kontinuiranih statističkih obilježaja X i Y.

Odgovarajuća f.r.v. može se zapisati u obliku

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, \mathrm{d}u\, \mathrm{d}v, \quad (x, y) \in \mathbf{R}^2, \tag{15}$$

pa se vidi da vrijedi

$$\frac{\partial^2 F(x, y)}{\partial x^2 \partial y^2} = f(x, y), \tag{16}$$

što za broj $f(x, y)$ opravdava naziv gustoća vjerojatnosti u točki $(x, y) \in \mathbf{R}^2$.

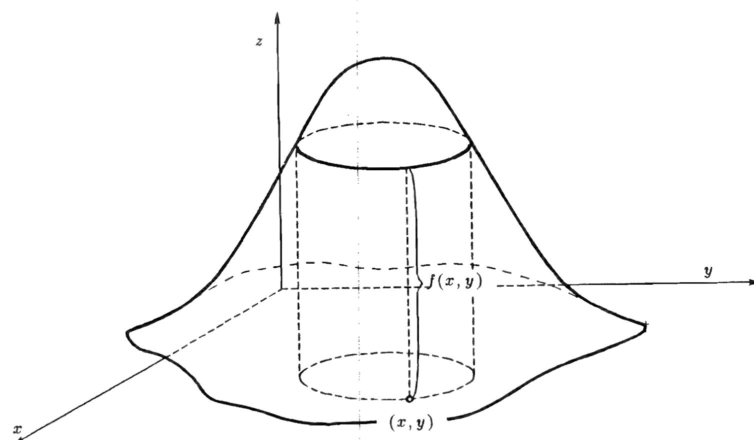
Ako je $S \subseteq \mathbf{R}^2$, onda se vjerojatnost da kontinuirani s.vk. (X, Y) poprimi vrijednost iz zadanog skupa S može izraziti formulom

$$\mathrm{P}(S) = \mathrm{P}((X, Y) \in S) = \iint_{(S)} f(x, y)\, \mathrm{d}x\, \mathrm{d}y, \tag{17}$$

tj. dobiva se integriranjem gustoća vjerojatnosti po skupu S.

Graf f.g.v. f, tj. skup $\{(x, y, z) \in \mathbf{R}^3 : z = f(x, y),\ (x, y) \in \mathbf{R}^2\}$, općenito će predstavljati neku plohu u prostoru. Ona se zove *ploha razdiobe*.

Najvažnija teorijska kontinuirana dvodimenzionalna razdioba vjerojatnosti je *dvodimenzionalna normalna (Gaussova) razdioba* s parametrima μ_1, μ_2, σ_1^2, σ_2^2 i ρ ($\sigma_1 > 0, \sigma_2 > 0, 0 \leq |\rho| < 1$). Piše se $(X, Y) \sim \mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ i govori da s.vk.



Slika 14. Skica plohe razdiobe

(X, Y) ima normalnu razdiobu s navedenim parametrima, ako pripadna f.g.v. glasi

(18) $$f(x,y) = K \exp[-Q(x,y)],$$

gdje je

(19) $$K = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}},$$

(20) $$Q(x,y) = \frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

K je, dakle, određena konstanta, ovisna o parametrima σ_1, σ_2 i ϱ, dok je $Q(x,y)$ pozitivno definitna kvadratna forma u varijablama x i y. Graf funkcije (18) zvonolika je ploha, čiji su presjeci s ravninama $z = c$ $(0 < c < K)$ elipse s jednadžbom oblika $Q(x,y) = c$. Elipse imaju zajedničko središte u točki (μ_1, μ_2) i toj točki pripada najveća gustoća vjerojatnosti $f(\mu_1, \mu_2) = K$, dok se udaljavanjem od te točke gustoća vjerojatnosti smanjuje.

Ako je (X, Y) kontinuirani s.vk., onda su pripadne marginalne razdiobe također kontinuirane vjerojatnosne razdiobe i komponenti X pripada funkcija gustoće vjerojatnosti

(21) $$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}\,y, \quad x \in \mathbf{R},$$

a komponenti Y funkcija gustoće vjerojatnosti

(22) $$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}\,x, \quad y \in \mathbf{R}.$$

Nadalje je

(23) $$\mu_1 = \mathrm{E}[X] = \int_{-\infty}^{\infty} x f_1(x)\,\mathrm{d}\,x, \quad \sigma_1^2 = \mathrm{V}[X] = \int_{-\infty}^{\infty} (x-\mu_1)^2 f_1(x)\,\mathrm{d}\,x,$$

(24) $$\mu_2 = \mathrm{E}[Y] = \int_{-\infty}^{\infty} y f_2(y)\,\mathrm{d}\,y, \quad \sigma_2^2 = \mathrm{V}[Y] = \int_{-\infty}^{\infty} (y-\mu_2)^2 f_2(y)\,\mathrm{d}\,y.$$

Uvjet nezavisnosti slučajnih varijabli X i Y sada se može zapisati u obliku

(25) $$f(x,y) = f_1(x) f_2(y), \quad (x,y) \in \mathbf{R}^2.$$

Pokazuje se da iz $(X, Y) \sim \mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ proizlazi da $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ i da $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, što znači da je $\mathrm{E}[X] = \mu_1$, $\mathrm{V}[X] = \sigma_1^2$, $\mathrm{E}[Y] = \mu_2$ i $\mathrm{V}[Y] = \sigma_2^2$. Za $\rho = 0$ vrijedi $f(x,y) = f_1(x)f_2(y)$, tj. tada su X i Y nezavisne slučajne varijable.

Ako je u dvodimenzionalnoj kontinuiranoj vjerojatnosnoj razdiobi $f_2(y) > 0$, onda se jednodimenzionalna kontinuirana razdioba sa funkcijom gustoće vjerojatnosti definiranom formulom

(26) $$p_y(x) = \frac{f(x,y)}{f_2(y)}, \quad x \in \mathbf{R},$$

zove *uvjetna razdioba vjerojatnosti* komponente X uz fiksiranu vrijednost y komponente Y. Veličina definirana formulom

(27) $$\mathrm{E}[X/y] = \int_{-\infty}^{\infty} x p_y(x)\,\mathrm{d}\,x,$$

zove se *uvjetno očekivanje* komponente X uz fiksiranu vrijednost y komponente Y.

Analogno se definiraju veličine

(28) $$q_x(y) = \frac{f(x,y)}{f_1(x)}, \quad y \in \mathbf{R},$$

(29) $$\mathrm{E}[Y/x] = \int_{-\infty}^{\infty} y q_x(y)\,\mathrm{d}\,y.$$

Tako se, na primjer, može pokazati (v. zad. 9) da je uvjetna razdioba komponente X za fiksiranu vrijednost y komponente Y u dvodimenzionalnoj normalnoj razdiobi $\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ normalna razdioba $\mathrm{N}(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2), \sigma_1^2(1-\rho^2))$, tako da je $\mathrm{E}[X/y] = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)$, dok je uvjetna razdioba komponente Y za fiksiranu vrijednost x komponente X normalna razdioba $\mathrm{N}(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1), \sigma_2^2(1-\rho^2))$, što znači da je $\mathrm{E}[Y/x] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)$.

4. Korelacija

Stavi li se $E[XY] = \sum_{i=1}^{r} \sum_{j=1}^{s} a_i a_j p_{ij}$ za diskretnu, odnosno $E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x,y)\, \mathrm{d}x\, \mathrm{d}y$ za kontinuiranu dvodimenzionalnu vjerojatnosnu razdiobu, može se reći da je $E[XY]$ očekivanje produkta slučajnih varijabli X i Y. Ako su X i Y nezavisne s.v., onda je $E[XY] = E[X]E[Y]$. Općenito se parametar μ_{11} ili $\mathrm{Cov}(X,Y)$, definiran formulom

(30) $$\mu_{11} = \mathrm{Cov}(X,Y) = E[XY] - E[X]E[Y],$$

zove ***korelacijski moment*** ili ***kovarijanca*** slučajnih varijabli X i Y.

Ako je $\mu_{11} = 0$, onda se kaže da su X i Y ***nekorelirane*** s.v. Nezavisne s.v. su, dakako, i nekorelirane, ali obratno općenito ne vrijedi.

Ako je $V[X] > 0$ i $V[Y] > 0$, onda se parametar

(31) $$\rho = \frac{\mu_{11}}{\sigma_1 \sigma_2}$$

zove ***koeficijent korelacije*** slučajnih varijabli X i Y.

Može se dokazati (v. zad. 10c) da je $\rho^2 \le 1$ i $\rho^2 = 1$ onda i samo onda ako između slučajnih varijabli X i Y postoji funkcijska ovisnost oblika $AX + BY + C = 0$ ($A \ne 0$ i $B \ne 0$). Značenje teorijskog koeficijenta korelacije definiranog u (31) slično je značenju empirijskog koeficijenta korelacije koje je opisano u III.5.

Pojam uvjetnog očekivanja, definiran u (12) i (14) za diskretnu, a u (27) i (29) za kontinuiranu dvodimenzionalnu razdiobu vjerojatnosti, omogućuje da se definiraju *funkcije regresije*

(32) $$x \mapsto E[Y/x], \quad y \mapsto E[X/y].$$

Funkcija $x \mapsto E[Y/x]$ pokazuje ovisnost uvjetnog očekivanja komponente Y o vrijednosti x komponente X, a slično značenje ima i druga funkcija regresije.

Graf funkcije regresije zove se *krivulja regresije*.

Ako je riječ o nezavisnim slučajnim varijablama X i Y, onda uvjetna razdioba jedne komponente ne ovisi o odabranoj vrijednosti druge komponente i jednaka je odgovarajućoj marginalnoj razdiobi (v. zad. 11), tako da je tada $E[Y/x] = E[Y]$ i $E[X/y] = E[X]$, što znači da su funkcije regresije konstante. Krivulje regresije su tada pravci usporedni s koordinatnim osima, koji se sijeku u točki (μ_1, μ_2).

Pravci koji, u smislu metode najmanjih kvadrata, najbolje aproksimiraju krivulje regresije zovu se *pravci regresije*. Njihove su jednadžbe

(33) $$y - \mu_2 = \frac{\mu_{11}}{\sigma_1^2}(x - \mu_1), \quad x - \mu_1 = \frac{\mu_{11}}{\sigma_2^2}(y - \mu_2),$$

iz čega se vidi da se oni sijeku u točki (μ_1, μ_2), a za kut φ između njih vrijedi

(34) $$\mathrm{tg}\,\varphi = \frac{1 - \rho^2}{\rho} \frac{\sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2}.$$

Za $\rho^2 = 1$ pravci regresije se međusobno poklapaju i tada se, kaže da je dvodimenzionalna *razdioba vjerojatnosti degenerirana*, tj. radi se zapravo o razdiobi vjerojatnosti na pravcu $y - \mu_2 = \frac{\sigma_2}{\sigma_1}(x - \mu_1)$.

Iz (31) se vidi da su X i Y, za $\rho = 0$, nekorelirane s.v., a iz (33) i (34) se razabire da su tada pravci regresije međusobno okomiti i usporedni s koordinatnim osima.

Za trinomnu razdiobu $B(m, p_1, p_2)$, na primjer, funkcije regresije su

(35) $$i \mapsto E[Y/i] = \frac{p_2}{1 - p_1}(m - i), \quad i = 0, 1, \ldots, m,$$

(36) $$j \mapsto E[X/j] = \frac{p_1}{1 - p_2}(m - j), \quad j = 0, 1, \ldots, m.$$

Budući da je riječ o diskretnoj vjerojatnosnoj razdiobi, krivulje regresije sastoje se iz diskretnog skupa točaka koje pripadaju istom pravcu. Prema tome, pravci regresije imaju jednadžbe

(37) $$y = \frac{p_2}{1 - p_1}(m - x), \quad x = \frac{p_1}{1 - p_2}(m - y).$$

Korelacijski je moment $\mu_{11} = -mp_1p_2$, a odgovarajući koeficijent korelacije $\rho = -\sqrt{\frac{p_1 p_2}{(1 - p_1)(1 - p_2)}}$, pa se vidi da se, za $p_1 + p_2 = 1$, dobiva $\rho = -1$, što znači da je dvodimenzionalana vjerojatnosna razdioba degenerirala u jednodimenzionalnu vjerojatnosnu razdiobu duž pravca $x + y = m$.

Za dvodimenzionalnu normalnu razdiobu $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ korelacijski je moment $\mu_{11} = \rho\sigma_1\sigma_2$, tako da parametar ρ ima značenje koeficijenta korelacije. Pravci regresije $y - \mu_2 = \rho\frac{\sigma_1}{\sigma_2}(x - \mu_1)$ i $x - \mu_1 = \rho\frac{\sigma_2}{\sigma_1}(y - \mu_2)$ ujedno su i krivulje regresije, tako da je ovdje riječ o *linearnoj regresiji*. Sada nekoreliranost ($\rho = 0$) implicira i nezavisnost s.v. X i Y, pa se može reći da su nezavisnost i nekoreliranost u dvodimenzionalnoj normalnoj razdiobi ekvivalentna svojstva, dok općenito nisu.

5. Višedimenzionalna razdioba vjerojatnosti

Matematički model za simultano promatranje n ($n \ge 2$) statističkih obilježja je slučajni vektor $(X_1, \ldots, X_n)$, čije komponente su s.v. $X_1, \ldots, X_n$. Svaki rezultat simultanog mjerenja n veličina uređena je n-torka $(x_1, \ldots, x_n) \in \mathbf{R}^n$, koja se zove *vrijednost slučajnog vektora* $(X_1, \ldots, X_n)$. Statistička zakonitost izražava se odgovarajućom funkcijom razdiobe vjerojatnosti na skupu $\mathbf{R}^n$, koja se definira formulom

(38) $$F(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n), \quad (x_1, \ldots, x_n) \in \mathbf{R}^n,$$

pa se govori da je formulom (38) definirana f.r.v. slučajnog vektora $(X_1, \ldots, X_n)$. To je realna funkcija n realnih varijabli i broj $F(x_1, \ldots, x_n)$ označava teorijsku vjerojatnost da se pri simultanom mjerenju slučajnih varijabli $X_1, \ldots, X_n$ dobije za X_1 vrijednost koja ne premašuje realni broj x_1, za X_2 vrijednost koja ne premašuje x_2 itd.

Funkcija $x_i \mapsto F_i(x_i) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty) = \mathrm{P}(X_i \leq x_i)$, $x_i \in \mathbf{R}$, jest f.r.v. jednodimenzionalne razdiobe vjerojatnosti koja se zove *marginalna razdioba vjerojatnosti komponente* X_i $(i = 1, \ldots, n)$. Ona opisuje statističko ponašanje (zakonitost) s.v. X_i same za sebe.

Ako vrijedi

(39) $$F(x_1, \ldots, x_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n),$$

za svako $(x_1, \ldots, x_n) \in \mathbf{R}^n$, onda se kaže da su $X_1, \ldots, X_n$ ***nezavisne slučajne varijable***.

Uređena n-torka $(\mu_1, \ldots, \mu_n) \in \mathbf{R}^n$, gdje je $\mu_i = \mathrm{E}[X_i]$ $(i = 1, \ldots, n)$, zove se ***vektor očekivanja*** slučajnog vektora $(X_1, \ldots, X_n)$, odnosno *središte* n-dimenzionalne razdiobe vjerojatnosti.

Ako je $n > 2$ onda se mogu promatrati i *marginalne dvodimenzionlne razdiobe vjerojatnosti* danoga slučajnog vektora $(X_1, \ldots, X_n)$. Uzme li se, naime, $i < j$ $(i, j = 1, \ldots, n)$ i definira funkcija

(40) $$F_{ij}(x_i, x_j) = F(\infty, \ldots, \infty, x_i, \infty, \ldots, \infty, x_j, \infty, \ldots, \infty) = \mathrm{P}(X_i \leq x_i, X_j \leq x_j), \quad (x_i, x_j) \in \mathbf{R}^2,$$

vidi se da je to f.r.v. slučajnog vektora (X_i, X_j). To omogućuje da se definira kvadratna matrica $\boldsymbol{\Sigma}$, čiji su elementi

(41) $$\sigma_{ij} = \mathrm{E}[X_i X_j] - \mathrm{E}[X_i]\mathrm{E}[X_j], \quad i, j = 1, \ldots, n,$$

i koja se zove ***kovarijancna*** ili ***disperzijska matrica***. To je simetrična matrica koja za dijagonalne elemente ima varijance $\sigma_{ii} = \sigma_i^2 = \mathrm{V}[X_i]$ $(i = 1, \ldots, n)$, dok su izvandijagonalni elementi kovarijance $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ $(i \neq j, \;\; i, j = 1, \ldots, n)$. Ako je $\sigma_i > 0$ $(i = 1, \ldots, n)$, onda se može definirati i kvadratna matrica $\boldsymbol{P}$ s elementima

(42) $$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \quad i, j = 1, \ldots, n,$$

koja se zove ***korelacijska matrica***. Dijagonalni elementi su jedinice $(\rho_{ii} = 1)$, a izvandijagonalni element ρ_{ij} $(i \neq j)$ označuje koeficijent korelacije slučajnih varijabli X_i i X_j.

Korelacijska matrica također je simetrična matrica i vrijedi

$$\det \boldsymbol{P} = \frac{1}{\sigma_1^2 \cdot \ldots \cdot \sigma_n^2} \det \boldsymbol{\Sigma},$$

tako da korelacijska i kovarijancna matrica imaju isti *rang* r $(r \leq n)$, pa se govori da n-dimenzionalna vjerojatnosna razdioba ima rang r.

Ako je $r < n$, kaže se da je n-dimenzionalna vjerojatnosna razdioba *degenerirana*.

Ako su $X_1, \ldots, X_n$ nezavisne s.v., onda je kovarijancna matrica dijagonalna, a korelacijska matrica jedinična matrica. Ako je, pak, $\boldsymbol{\Sigma}$ dijagonalna matrica, onda

se kaže da su $X_1, \ldots, X_n$ ***nekorelirane slučajne varijable***. Nezavisne slučajne varijable su, dakle, nekorelirane, dok obratno općenito ne vrijedi.

Najvažniji primjer teorijskog modela n-dimenzionalne vjerojatnosne razdiobe je ***normalna (Gaussova) n-dimenzionalna razdioba***. Kaže se da s.vk. $(X_1, \ldots, X_n)$ ima normalnu razdiobu s vektorom očekivanja $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \in \mathbf{R}^n$ i kovarijancnom matricom $\boldsymbol{\Sigma}$ ($\boldsymbol{\Sigma}$ je regularna i pozitivno definitna) i piše se $(X_1, \ldots, X_n) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, ako se pripadna f.r.v. može zapisati u obliku

(43) $$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n) \, \mathrm{d}t_1 \ldots \mathrm{d}t_n,$$

gdje je

(44) $$f(t_1, \ldots, t_n) = K \exp[-Q(t_1, \ldots, t_n)], \quad (t_1, \ldots, t_n) \in \mathbf{R}^n,$$

(45) $$K = (2\pi \det \boldsymbol{\Sigma})^{-\frac{n}{2}},$$

(46) $$Q(t_1, \ldots, t_n) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{ij}(t_i - \mu_i)(t_j - \mu_j),$$

a λ_{ij} su elementi matrice $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ ($\boldsymbol{\Sigma}^{-1}$ je inverzna matrica od $\boldsymbol{\Sigma}$). Formulom (44) definirana je f.g.v. n-dimenzionalne normalne razdiobe $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Formulom (46) definirana je, pak, određena pozitivno definitna kvadratna forma u varijablama $t_1, \ldots, t_n$.

Može se dokazati (v. [6]) da komponenti X_i pripada marginalna razdioba $\mathrm{N}(\mu_i, \sigma_i^2)$ $(i = 1, \ldots, n)$, iz čega slijedi da će $X_1, \ldots, X_n$ biti nezavisne s.v. onda i samo onda ako je $\boldsymbol{\Sigma}$, pa dakle i $\boldsymbol{\Lambda}$, dijagonalna matrica, tj. ako su $X_1, \ldots, X_n$ nekorelirane slučajne varijable.

6. Funkcije više slučajnih varijabli

Ako je $(x_1, \ldots, x_n) \mapsto h(x_1, \ldots, x_n)$ određena realna funkcija n $(n \geq 2)$ realnih varijabli i ako je $(X_1, \ldots, X_n)$ zadani slučajni vektor, onda je $Y = h(X_1, \ldots, X_n)$ s.v. za koju se kaže da je *funkcija slučajnog vektora* $(X_1, \ldots, X_n)$.

Ako je, na primjer, $h(x_1, \ldots, x_n) = a_1 x_1 + \ldots + a_n x_n$, gdje su $a_1, \ldots, a_n$ zadani realni brojevi (koeficijenti), onda se kaže da je $Y = a_1 X_1 + \ldots + a_n X_n$ *linearna kombinacija slučajnih varijabli* $X_1, \ldots, X_n$. Može se dokazati (v. zad. 16) da vrijedi

(47) $$\mathrm{E}[Y] = \mathrm{E}[a_1 X_1 + \ldots + a_n X_n] = a_1 \mathrm{E}[X_1] + \ldots + a_n \mathrm{E}[X_n],$$

(48) $$\mathrm{V}[Y] = \mathrm{V}[a_1 X_1 + \ldots + a_n X_n] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \sigma_{ij}.$$

Iz formule (47) vidi se da je očekivanje linearne kombinacije slučajnih varijabli jednako linearnoj kombinaciji očekivanja komponenata, dok se iz (48) razabire da

je varijanca linearne kombinacije slučajnih varijabli jednaka linearnoj kombinaciji elemenata kovarijancne matrice zadanoga slučajnog vektora.

Ako su $X_1, \ldots, X_n$ nekorelirane slučajne varijable, onda (48) postaje

(49) $$V[a_1X_1 + \ldots + a_nX_n] = a_1^2 V[X_1] + \ldots + a_n^2 V[X_n],$$

tj. tada je varijanca linearne kombinacije s.v. jednaka linearnoj kombinaciji varijanci komponenata s kvadriranim koeficijentima.

Glavni i najteži problemi u vezi s funkcijama slučajnih varijabli sastoje se u određivanju vjerojatnosne razdiobe slučajne varijable $Y = h(X_1, \ldots, X_n)$ uz zadanu funkciju h i vjerojatnosnu razdiobu slučajnog vektora $(X_1, \ldots, X_n)$.

Budući da se velik dio teorije statističkog zaključivanja (treći dio ove knjige) temelji baš na rješenjima spomenutih problema, čije izlaganje prelazi zamišljene okvire ove knjige, ovdje će se navesti neki najvažniji rezultati. Svi navedeni rezultati vrijede uz zajedničku pretpostavku da su $X_1, \ldots, X_n$ nezavisne slučajne varijable i to se u nastavku više neće isticati.

1. Ako $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, onda $Y = a_1X_1 + \ldots + a_nX_n \sim N(\mu, \sigma^2)$, gdje je $\mu = a_1\mu_1 + \ldots + a_n\mu_n$ i $\sigma^2 = a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2$.

Pojednostavnjeno, može se reći da je linearna kombinacija nezavisnih normalnih slučajnih varijabla također normalna s.v.

2. Ako $X_i \sim B(m_i, p)$, $i = 1, \ldots, n$, onda $Y = X_1 + \ldots + X_n \sim B(m, p)$, gdje je $m = m_1 + \ldots + m_n$.

To znači da je zbroj nezavisnih binomnih slučajnih varijabli zajedničkog parametra p također binomna s.v.

3. Ako $X_i \sim Po(\lambda_i)$, $i = 1, \ldots, n$, onda $Y = X_1 + \ldots + X_n \sim Po(\lambda)$, gdje je $\lambda = \lambda_1 + \ldots + \lambda_n$.

Zbroj nezavisnih Poissonovih slučajnih varijabli također je Poissonova slučajna varijabla.

4. Ako $X_i \sim G(\alpha, \beta_i)$, $i = 1, \ldots, n$, onda $Y = X_1 + \ldots + X_n \sim G(\alpha, \beta)$, gdje je $\beta = \beta_1 + \ldots + \beta_n$. Posebno, ako $X_i \sim G(\alpha, 1) = Ex(\alpha)$, onda $Y \sim G(\alpha, n)$. Ako, pak, $X_i \sim G\left(\frac{1}{2}, \frac{n_i}{2}\right) = \chi^2(n_i)$ $(n_i \in \mathbf{N})$, onda $Y \sim G\left(\frac{1}{2}, \frac{n}{2}\right) = \chi^2(n)$, gdje je $n = n_1 + \ldots + n_n$.

Može se, dakle, reći da je zbroj n nezavisnih eksponencijalnih slučajnih varijabli zajedničkog parametra α s.v. gama-razdiobe $G(\alpha, n)$, dok je zbroj nezavisnih slučajnih varijabli hikvadrat-razdiobe također s.v. hikvadrat-razdiobe.

5. Ako $X_i \sim N(0,1)$, $i = 1, \ldots, n$, onda $Y = X_1^2 + \ldots + X_n^2 \sim \chi^2(n)$.

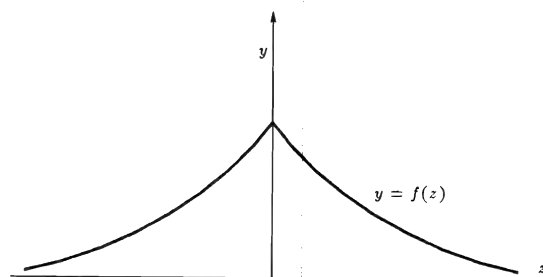
To pokazuje da zbroj kvadrata n nezavisnih standardnih normalnih slučajnih varijabli ima hikvadrat-razdiobu sa n stupnjeva slobode.

6. Ako su X i Y nezavisne s.v. i obje imaju eksponencijalnu razdiobu $Ex(\alpha)$, onda s.v. $Z = X - Y$ ima tzv. *Laplaceovu* ili *dvostruko eksponencijalnu razdiobu* parametra α $(\alpha > 0)$.

Pripadna f.g.v. glasi

$$f(z) = \frac{\alpha}{2}\exp(-\alpha|z|), \quad z \in \mathbf{R},$$

a pokazuje se da je $E[Z] = 0$ i $V[Z] = \frac{2}{\alpha^2}$.



Slika 15. Skica krivulje Laplaceove razdiobe

7. Ako su X i Y nezavisne s.v., $X \sim N(0,1)$ i $Y \sim \chi^2(n)$, onda slučajnoj varijabli $Z = X\sqrt{\frac{n}{Y}}$ pripada tzv. ***Studentova razdioba*** ili ***t-razdioba sa n stupnjeva slobode***, što se piše $Z \sim t(n)$.

Pripadna f.g.v. glasi

$$f(z) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{z^2}{n}\right)^{-\frac{n+1}{2}}, \quad z \in \mathbf{R}.$$

Za $n = 1$ dobiva se tzv. *Cauchyjeva razdioba* koja je zanimljiva zbog toga što nema konačno očekivanje, pa ni bilo koji moment višeg reda. Za Studentovu razdiobu $t(n)$ najvažniji parametri izraženi su formulama

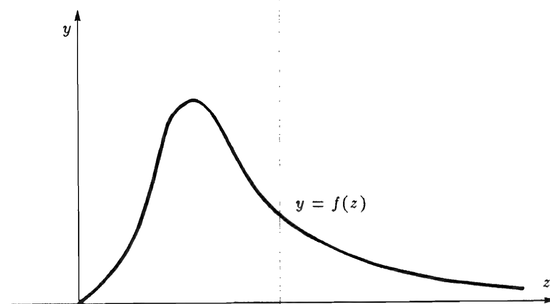
$$E[Z] = 0, \quad \text{za } n > 1,$$
$$V[Z] = \frac{n}{n-2}, \quad \text{za } n > 2.$$

Krivulja Studentove razdiobe slična je krivulji normalne razdiobe $\mathrm{N}(0,\sigma^2)$ (v. sl. 22. u VIII.6).

> 8. Ako su X i Y nezavisne s.v. i $X \sim \chi^2(r)$, a $Y \sim \chi^2(s)$ $(r, s \in \mathbf{N})$, onda s.v. $Z = \dfrac{sX}{rY}$ ima tzv. ***F-razdiobu*** *sa* (r,s) *stupnjeva slobode.*

Piše se $Z \sim \mathrm{F}(r,s)$, a pripadna f.g.v. glasi

$$f(z) = \begin{cases} 0, & \text{za } z \le 0 \\ \dfrac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)} r^{\frac{r}{2}} s^{\frac{s}{2}} \dfrac{z^{\frac{r}{2}-1}}{(rz+s)^{\frac{r+s}{2}}}, & \text{za } z > 0. \end{cases}$$



Slika 16. Skica tipične krivulje F-razdiobe

Očekivanje postoji za $s > 2$, a varijanca za $s > 4$ i vrijedi

$$\mathrm{E}[Z] = \frac{s}{s-2}, \qquad \mathrm{V}[Z] = \frac{2s^2(r+s-2)}{r(s-2)^2(s-4)}.$$

> 9. Ako su $X_1, \ldots, X_n$ nezavisne s.v. sa zajedničkom f. r. v. F, onda slučajnoj varijabli $Y = \max(X_1, \ldots, X_n)$ pripada funkcija razdiobe vjerojatnosti
>
> (50) $\qquad \mathrm{G}(y) = \mathrm{P}(Y \le y) = [\mathrm{F}(y)]^n, \quad y \in \mathbf{R},$
>
> a slučajnoj varijabli $Z = \min(X_1, \ldots, X_n)$ pripada funkcija razdiobe vjerojatnosti
>
> (51) $\qquad \mathrm{H}(z) = \mathrm{P}(Z \le z) = 1 - [1 - F(z)]^n, \quad z \in \mathbf{R}.$

Jasno je da se slučajni vektor $\boldsymbol{X} = (X_1, \ldots, X_n)$ može podvrći određenoj transformaciji (operatoru) tako da se kao rezultat opet dobije slučajni vektor, recimo $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$.

Posebno su važne *linearne transformacije* (*linearni operatori*), koje se mogu opisati odgovarajućim realnim matricama. Ako je, naime, $\boldsymbol{A}$ realna matrica tipa $m \times n$ (m je broj redaka, a n broj stupaca matrice $\boldsymbol{A}$) i ako se vektori $\boldsymbol{X}$ i $\boldsymbol{Y}$ tretiraju kao jednoredne matrice (tipa $1 \times n$, odnosno $1 \times m$), onda se djelovanje linearne transformacije može izraziti matričnom jednadžbom $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}^{\mathsf{T}}$, gdje $\boldsymbol{A}^{\mathsf{T}}$ označuje transponiranu matricu od $\boldsymbol{A}$.

> 10. Ako slučajni vektor $\boldsymbol{X} = (X_1, \ldots, X_n) \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}^{\mathsf{T}}$, gdje je $\boldsymbol{A}$ realna matrica tipa $m \times n$, onda za slučajni vektor $\boldsymbol{Y}$ vrijedi
>
> (52) $\qquad \boldsymbol{Y} = (Y_1, \ldots, Y_m) \sim \mathrm{N}(\boldsymbol{\mu}\boldsymbol{A}^{\mathsf{T}}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\mathsf{T}}).$

Zadaci

1. Ako je A skup točaka ravnine ($A \subseteq \mathbf{R}^2$), čija površina iznosi a ($a > 0$) mjernih jedinica (recimo m^2) i ako je funkcija f zadana formulom

$$f(x,y) = \begin{cases} \dfrac{1}{a}, & \text{za } (x,y) \in A \\ 0, & \text{za } (x,y) \notin A, \end{cases}$$

onda se govori o *jednolikoj* (*uniformnoj*) *razdiobi vjerojatnosti na skupu* A. Napišite pripadnu f. r. v. ako je A jedinični kvadrat.

2. Da li su X i Y nezavisne s.v. ako je riječ o uniformnoj razdiobi na:

a) jediničnom kvadratu,

b) jediničnom krugu,

c) trokutu s vrhovima u točkama (0,0), (1,0) i (0,1)?

3. Dokažite da su marginalne razdiobe za trinomnu razdiobu $\mathrm{B}(m, p_1, p_2)$ binomne razdiobe $\mathrm{B}(m, p_1)$ i $\mathrm{B}(m, p_2)$.

4. Dokažite da za veličine definirane u (11) i (13) vrijedi

$$\sum_{i=1}^{r} p_{i/j} = \sum_{j=1}^{s} q_{j/i} = 1.$$

5. Dokažite da su uvjetne razdiobe za trinomnu razdiobu $\mathrm{B}(m, p_1, p_2)$ binomne razdiobe $\mathrm{B}\left(m - j, \dfrac{p_1}{1-p_2}\right)$ i $\mathrm{B}\left(m - i, \dfrac{p_2}{1-p_1}\right)$.

6. Dokažite da za funkcije f_1 i f_2 definirane u (21) i (22) vrijedi

$$\int_{-\infty}^{\infty} f_1(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} f_2(y)\,\mathrm{d}y = 1.$$

7. Dokažite da iz $F(x,y) = F_1(x)F_2(y)$ proizlazi formula (25).

8. Dokažite da za veličine definirane formulama (26) i (28) vrijedi

$$\int_{-\infty}^{\infty} p_y(x)\,\mathrm{d}\,x = \int_{-\infty}^{\infty} q_x(y)\,\mathrm{d}\,y = 1.$$

9. Dokažite da je uvjetna razdioba komponente X u dvodimenzionalnoj normalnoj razdiobi $\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ normalna razdioba $\mathrm{N}(\,\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2),$ $\sigma_1^2(1-\rho^2)\,)$.

10. Dokažite da za diskretne i kontinuirane s.v. X i Y vrijedi:

a) ako su X i Y nezavisne, onda je $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$,

b) ako su X i Y nezavisne, onda su i nekorelirane s.v.,

c) $\rho^2 \leq 1$, gdje je ρ koeficijent korelacije.

11. Dokažite da vrijedi:

a) $p_{i/j} = q_j$ i $q_{j/i} = p_i$, ako su X i Y nezavisne diskretne slučajne varijable,

b) $p_y(x) = f_1(x)$ i $q_x(y) = f_2(y)$, ako su X i Y nezavisne kontinuirane slučajne varijable.

12. Izvedite jednadžbe pravaca regresije izražene formulama (33) .

13. Dokažite da za kut φ između pravaca regresije vrijedi formula (34).

14. Za trinomnu razdiobu izvedite formule za:

a) vjerojatnosti uvjetnih razdioba,

b) funkcije regresije.

15. Za dvodimenzionalnu normalnu razdiobu izvedite formule za:

a) uvjetne f.g.v.,

b) funkcije regresije.

16. Dokažite da za diskretne i kontinuirane dvodimenzionalne vjerojatnosne razdiobe vrijedi:

a) $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$,

b) $\mathrm{V}[aX + bY] = a^2\mathrm{V}[X] + b^2\mathrm{V}[Y] + 2ab\,\mathrm{Cov}(X, Y)$, gdje su a i b proizvoljni realni brojevi.

17. Dokažite da je očekivanje Laplaceove razdiobe parametra $\alpha > 0$ nula, a varijanca $\frac{2}{\alpha^2}$.

18. Izvedite formule (50) i (51).

19. Odredite konstantu c $(c > 0)$ tako da vrijedi $\mathrm{P}(|X| \geq c) = \alpha$ $(\alpha = 0{,}10\,;\ 0{,}05\,;\ 0{,}01)$ ako:

a) $X \sim \mathrm{N}(0, 1)$, b) $X \sim \mathrm{U}(-\sqrt{3}, \sqrt{3})$, c) $X \sim \mathrm{t}(1)$,

d) X ima Laplaceovu razdiobu parametra $\alpha = \sqrt{2}$.

Skicirajte u istom pravokutnome koordinatnom sustavu odgovarajuće krivulje razdiobe i zatim geometrijski interpretirajte dobivene rezultate.

20. Odredite konstante c_1 i c_2 $(c_1 > 0, c_2 > 0)$ tako da vrijedi $\mathrm{P}(X \leq c_1) = \mathrm{P}(X \geq c_2) = \frac{\alpha}{2}$ $(\alpha = 0{,}10\,;\ 0{,}05\,;\ 0{,}01)$ ako:

a) $X \sim \mathrm{N}(32, 64)$, b) $X \sim \chi^2(20)$, c) $X \sim \mathrm{F}(5, 8)$.

Uputa: Poslužite se tabl. 3, 4. i 7. u Dodatku.

21. Neka je (X, Y) slučajni vektor s kovarijancom $\mathrm{Cov}(X, Y)$ i α_0, α_1, β_0 i β_1 realni brojevi, te $X_1 = \alpha_1 X + \alpha_0$ i $Y_1 = \beta_1 Y + \beta_0$. Dokažite da je $\mathrm{Cov}(X_1, Y_1) = \alpha_1\beta_1\,\mathrm{Cov}(X, Y)$.

22. Neka je $\boldsymbol{X} = (X_1, \ldots, X_n)$ slučajni vektor s vektorom očekivanja $\mathrm{E}[\boldsymbol{X}] = (\mathrm{E}[X_1], \ldots, \mathrm{E}[X_n])$ i kovarijancnom matricom $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, te $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n) \in \mathbf{R}^n$ i $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{\lambda}$, gdje se $\boldsymbol{X}$, $\boldsymbol{Y}$ i $\boldsymbol{\lambda}$ tretiraju kao jednoredne matrice, a $\boldsymbol{X} + \boldsymbol{\lambda}$ kao zbroj matrica.
Dokažite da vrijedi:

a) $\mathrm{E}[\boldsymbol{Y}] = \mathrm{E}[\boldsymbol{X} + \boldsymbol{\lambda}] = \mathrm{E}[\boldsymbol{X}] + \boldsymbol{\lambda}$, gdje je $\mathrm{E}[\boldsymbol{Y}]$ vektor očekivanja slučajnog vektora $\boldsymbol{Y}$, a $\mathrm{E}[\boldsymbol{X}]$ i $\mathrm{E}[\boldsymbol{Y}]$ također se tretiraju kao jednoredne matrice,

b) $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, gdje je $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ kovarijancna matrica slučajnog vektora $\boldsymbol{Y}$.

Uputa: Primijenite formulu iz zad. 21.

23. Neka je $\boldsymbol{X} = (X_1, \ldots, X_n)$ slučajni vektor s vektorom očekivanja $\mathrm{E}[\boldsymbol{X}]$ i kovarijancnom matricom $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, te $\boldsymbol{A}$ realna matrica tipa $m \times n$ i $\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{A}^{\mathrm{T}}$.
Dokažite da vrijedi:

a) $\mathrm{E}[\boldsymbol{Y}] = \mathrm{E}[\boldsymbol{X}]\,\boldsymbol{A}$, gdje je $\mathrm{E}[\boldsymbol{Y}]$ vektor očekivanja slučajnog vektora $\boldsymbol{Y}$.

b) $\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{A}\,\boldsymbol{\Sigma}_{\boldsymbol{X}}\,\boldsymbol{A}^{\mathrm{T}}$, gdje je $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ kovarijancna matrica slučajnog vektora $\boldsymbol{Y}$.

24. Dokažite da se formula (48) iz V.6. može zapisati u obliku $\mathrm{V}[Y] = \mathrm{V}[\boldsymbol{a}\,\boldsymbol{X}] = \boldsymbol{a}\,\boldsymbol{\Sigma}_{\boldsymbol{X}}\,\boldsymbol{a}^{\mathrm{T}}$, gdje se vektor $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbf{R}^n$ također tretira kao jednoredna matrica.

Pregled važnijih teorijskih razdioba vjerojatnosti

Naziv	Oznaka	Očekivanje	Varijanca
binomna	$B(m,p)$ $m \in \mathbf{N}, 0 < p < 1$	mp	$mp(1-p)$
Poissonova	$Po(\lambda), \lambda > 0$	λ	λ
geometrijska	$0 < p < 1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
normalna	$N(\mu, \sigma^2)$ $\mu \in \mathbf{R}, \sigma > 0$	μ	σ^2
gama	$G(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$\frac{\beta}{\alpha}$	$\frac{\beta}{\alpha^2}$
eksponencijalna	$Ex(\alpha), \alpha > 0$	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$
hikvadrat	$\chi^2(n), n \in \mathbf{N}$	n	$2n$
beta	$\alpha > 0, \beta > 0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
uniformna	$U(a,b), a < b$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b^2-a^2)$
lognormalna	$LN(\mu, \sigma^2)$ $\mu \in \mathbf{R}, \sigma > 0$	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$	$\exp(2\mu + \sigma^2)(\exp \sigma^2 - 1)$
Laplaceova	$\alpha > 0$	0	$\frac{2}{\alpha^2}$
Studentova	$t(n), n \in \mathbf{N}$	$0 \ (n > 1)$	$\frac{n}{n-2} \ (n > 2)$
Cauchyjeva	$t(1)$	–	–
F-razdioba	$F(r,s)\ ;\ r, s \in \mathbf{N}$	$\frac{s}{s-2} \ (s > 2)$	$\frac{2s^2(r+s-2)}{r(s-2)^2(s-4)} \ (s > 4)$

TREĆI DIO

TEORIJA STATISTIČKOG ZAKLJUČIVANJA

Statistički fenomeni očituju se u statističkim podacima. Proučavanjem i analizom statističkih podataka onako kako je opisano u prvom dijelu došlo se do sinteze izražene u obliku teorije statističkih fenomena, ukratko prikazane u drugom dijelu. Glavnim problemom matematičke statistike smatra se ipak kako, na temelju konačnog broja statističkih podataka, izvesti korektne zaključke o promatranome statističkom fenomenu.

Općenito govoreći teorija statističkog zaključivanja proučava odnose između konačnog niza statističkih podataka i matematičkih modela izgrađenih u ***teoriji slučajnih varijabli.*** Istraživači statističkih pojava suočeni su sa zadatkom da otkrivaju statističke zakonitosti i pripadne parametre na temelju konačnog niza podataka, dok je za potpuno određenje te zakonitosti redovito nužno beskonačno mnogo statističkih podataka. Govori se da se zaključci donose na temelju konačnog ***uzorka.*** Stoga zaključci neće imati apsolutnu sigurnost, već će se govoriti o određenoj ***pouzdanosti*** izvedenog zaključka.

U teoriji statističkog zaključivanja izgrađuju se matematički modeli koji omogućuju egzaktno definiranje problema i njihovo rješavanje matematičkim metodama, a također i primjenu dobivenih rezultata u praktičnom životu i drugim znanstvenim disciplinama. Drugim riječima, za određene praktične situcije konstruiraju se odgovarajući ***teorijsko-statistički modeli*** i zatim se pronalaze ***statističk[e] metode*** kojima se postiže "zadovoljavajuće" rješenje.

Opća je pretpostavka pri izgradnji svih teorijskih modela za statistička zaključivanja da je dani niz $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ statističkih podataka vrijednost određenoga slučajnog vektora $(X_1, \ldots, X_n) = \boldsymbol{X}$. Osnovnu ulogu u izgradnji modela ima definiranje ***klase*** $\mathcal{P}$ svih ***dopuštenih vjerojatnosnih razdioba*** za slučajni vektor $\boldsymbol{X}$. Svaki se, naime, problem statističkog zaključivanja svodi na pitanje što se može reći o vjerojatnosnoj razdiobi slučajnog vektora $\boldsymbol{X}$ na temelju danog ***vektora podatak[a]*** $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbf{R}^n$. Za definiranje klase $\mathcal{P}$ ne postoje egzaktni teorijski kriteriji, već se to obično čini na temelju iskustva i intuicije. Ako se uzme preuska klasa $\mathcal{P}$ dopuštenih razdioba vjerojatnosti, onda postoji velika mogućnost da stvarna raz[-] dioba vjerojatnosti ostane izvan te klase, tj. izvan modela, a ako se za $\mathcal{P}$ uzm[e] preširoka klasa, recimo klasa svih n-dimenzionalnih vjerojatnosnih razdioba, ond[a] se praktički ništa ne može zaključiti o stvarnoj razdiobi P na temelju vektora p[o]dataka $\boldsymbol{x}$. Stoga se usvajaju ona ograničenja na dopuštene vjerojatnosne razdiob[e] koja se temelje na prirodi promatranog problema, te iskustvu i intuiciji istraživač[a].

Redovito se usvaja pretpostavka da su $X_1, \ldots, X_n$ nezavisne slučajne varijab[le] sa zajedničkom (jednodimenzionalnom) vjerojatnosnom razdiobom P. To odg[o]vara praktičnoj situaciji kada su $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ nezavisna mjerenja slučajne varijab[le] X, kojoj pripada f.r.v. $F(x) = P(X \leq x), x \in \mathbf{R}$, istraživaču redovito nepoznat[a]. Drugim riječima, pretpostavlja se da je rezultat svakog mjerenja posljedica jed[-]

te iste statističke zakonitosti. Time je klasa $\mathcal{P}$ sužena na sve moguće jednodimenzionalne vjerojatnosne razdiobe. No i to je, redovito, preširoka klasa dopuštenih razdioba. Stoga će se, na primjer, često klasa $\mathcal{P}$ definirati kao klasa svih mogućih normalnih razdioba, tj. uzimat će se da je $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma > 0\}$, pa će pitanje glasiti: što se može reći o parametrima μ i σ^2 na temelju danog niza podataka $x_1, \ldots, x_n$.

Općenito, ako se klasa $\mathcal{P}$ svih dopuštenih vjerojatnosnih razdioba može opisati pomoću konačnog broja parametara, onda se govori o *parametarskom modelu* statističkog zaključivanja.

Postoje određeni problemi statističkog zaključivanja, kao, na primjer, utvrđivanje nezavisnosti slučajnih varijabli X i Y na temelju niza mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$, u kojima se ne može klasa $\mathcal{P}$ opisati na jednostavan način pomoću konačnog broja parametara i tada se govori o *neparametarskom modelu*.

Tipični su problemi statističkog zaključivanja *problem procjene parametara*, koji se sastoji u pronalaženju numeričke vrijednosti kojom se aproksimira nepoznati parametar pretpostavljene vjerojatnosne razdiobe i određuje točnost te aproksimacije, te *problem testiranja hipoteze* u kojem se postavlja zadatak definiranja postupka za donošenje odluke o prihvaćanju, odnosno odbacivanju, unaprijed istaknute hipoteze o razdiobi vjerojatnosti.

Teorija statističkog zaključivanja zapravo se i sastoji od različitih modela i metoda koje su razvijene za rješavanje niza problema tipa procjene parametara i testiranja hipoteza, tako da se danas još ne može smatrati cjelovitom i konzistentnom teorijom. To je relativno "mlada" teorija, koja se brzo razvija, ali u kojoj još uvijek postoje važna pitanja na koja nema zadovoljavajućih odgovora.

VI. Procjena parametara

1. Uvod u problematiku

Radi lakšeg razumijevanja problema procjene parametara i općih pojmova koji se u vezi s tim definiraju, razmotrit će se jedan tipičan primjer.

1. primjer

Da bi se procijenio nepoznati parametar p, koji označuje proporciju ($100\,p$ je postotak) neispravnih proizvoda u određenome tehnološkom procesu u kojem se masovno proizvodi neki proizvod, ispitat će se n proizvoda i zabilježiti niz podataka $x_1, \ldots, x_n$, gdje je

$$x_i = \left\{ \begin{array}{ll} 1, & \text{ako je proizvod neispravan} \\ 0, & \text{ako proizvod nije neispravan} \end{array} \right\} \quad i = 1, \ldots, n.$$

Prvi korak u rješavanju problema procjene nepoznatog parametra p svakako je definiranje postupka (funkcije) kojim se iz danih podataka izračunava numerička vrijednost procjene. Intuitivno se čini razumnim pretpostaviti da je aritmetička sredina dobivenih podataka

$$\text{(1)} \qquad \overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

dobra procjena za nepoznati parametar p. Vidi se, naime, da je $\overline{x}$ zapravo relativna frekvencija, tj. proporcija neispravnih proizvoda u nizu od n ispitanih proizvoda. No, odmah se nameće i zadatak da se egzaktnije utvrdi zašto je $\overline{x}$ dobra procjena za p i, ako je moguće, da se i kvantitativno izrazi "kvaliteta" te procjene. Trebalo bi, naime, ustanoviti kakva je greška kada se p aproksimira sa $\overline{x}$ i kako bi se procjena (1) usporedila s nekom drugom procjenom za nepoznati parametar p.

Da bi se dobili odgovori na ta pitanja, treba definirati odgovarajući teorijski model. U tu svrhu pretpostavlja se da je $(x_1, \ldots, x_n)$ vrijednost slučajnog vektora $(X_1, \ldots, X_n)$, gdje su $X_1, \ldots, X_n$ nezavisne s.v. sa zajedničkom binomnom (Bernoullijevom) razdiobom $B(1, p)$, tj. vrijedi

$$\text{(2)} \qquad P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p, \quad i = 1, \ldots, n.$$

Pretpostavlja se, zapravo, da se proces proizvodnje u pogledu broja defektnih proizvoda pokorava statističkoj zakonitosti opisanoj binomnom razdiobom $B(1, p)$. Time je definirana klasa dopuštenih vjerojatnosnih razdioba

$$\mathcal{P} = \{B(1, p) : 0 < p < 1\}.$$

Empirijska veličina $\overline{x}$ može se tada shvatiti kao vrijednost slučajne varijable

(3) $$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Ako se, naime, zamisli višestruko ponavljanje ispitivanja serija od po n proizvoda i za svaki dobiveni niz podataka izračuna odgovarajuća aritmetička sredina, onda će se dobivene empirijske vrijednosti $\overline{x}$ pokoravati teorijskoj vjerojatnosnoj razdiobi slučajne varijable $\overline{X}$.

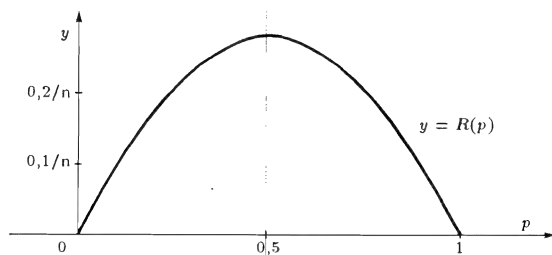
Slučajna varijabla $\overline{X}$, kao funkcija od slučajnih varijabli $X_1, \ldots, X_n$, ima vjerojatnosnu razdiobu ovisnu o nepoznatom parametru p i prirodnom broju n. Poznato je (v. IV.3. i V.6. točka 2) da u pretpostavljenome teorijskom modelu $X_1 + \ldots + X_n \sim \mathrm{B}(n, p)$, tako da je

(4) $$\mathrm{P}\left(\overline{X} = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Formula (4) omogućuje da se teorijski spozna kako se vrijednosti $\overline{x}$ ponašaju u odnosu na nepoznati parametar p. Iz (4) odmah proizlazi da je $\mathrm{E}[\overline{X}] = p$, što govori da su one raspršene oko p, kao svojega matematičkog očekivanja. Najprirodnija mjera raspršenja svakako je varijanca (v. II.3). Stoga se prirodno nameće ideja da se promotri funkcija

(5) $$p \mapsto R(p) = \mathrm{V}[\overline{X}] = \frac{1}{n} p(1-p), \quad p \in \langle 0, 1 \rangle,$$

koja omogućuje određeni uvid u točnost aproksimacije nepoznatog parametra p vrijednošću $\overline{x}$.



Slika 1. Graf funkcije $p \mapsto R(p)$

Budući da je $\mathrm{V}[\overline{X}] = E[(\overline{X} - p)^2]$, može se reći da $R(p) = \mathrm{E}[(\overline{X} - p)^2]$ pokazuje očekivanu kvadratnu grešku pri procjeni nepoznatog parametra p aritmetičkom sredinom $\overline{x}$ izmjerenih podataka. Slika 1. pokazuje da je ta greška maksimalna za $p = 0{,}5$ i tada iznosi $\frac{0{,}25}{n}$. To znači da pri procjeni nepoznatoga teorijskog parametra p empirijskom vrijednošću $\overline{x}$, dobivenom na temelju n nezavisnih mjerenja, očekivana kvadratna greška neće premašiti $\frac{0{,}25}{n}$.

Općenito, ako je riječ o parametarskom modelu, onda se klasa $\mathcal{P}$ svih dopuštenih vjerojatnosnih razdioba može izraziti zapisom
$$\mathcal{P} = \{P_t : t \in \Theta\},$$
gdje je $\Theta(\Theta \subseteq \mathbf{R}^k)$ skup svih dopuštenih vrijednosti parametara pretpostavljene vjerojatnosne razdiobe. Za $k=1$ imamo *jednoparametarski model*, a za $k>1$ ($k \in \mathbf{N}$) *višeparametarski model*. $P_t(\cdot)$ će označivati vjerojatnost dotičnog događaja uz pretpostavku da parametar razdiobe ima konkretnu vrijednost t.

Slučajni vektor $(X_1, \ldots, X_n)$, čije su komponente nezavisne slučajne varijable sa zajedničkom vjerojatnosnom razdiobom P_t, tj. čija f.r.v. ima oblik

(6) $$F(x_1, \ldots, x_n) = P_t(X_1 \le x_1) \cdot \ldots \cdot P_t(X_n \le x_n), \quad (x_1, \ldots, x_n) \in \mathbf{R}^n,$$

zove se ***slučajni uzorak** veličine n*.

Procjenitelj ili ***estimator*** nepoznatog parametra t je s.v. $\widehat{T}$, definirana kao određena funkcija slučajnog uzorka $(X_1, \ldots, X_n)$. Piše se

(7) $$\widehat{T} = h(X_1, \ldots, X_n),$$

gdje je $(x_1, \ldots, x_n) \mapsto h(x_1, \ldots, x_n)$ određena realna funkcija n realnih varijabli.

Budući da je u matematičkoj statistici uobičajeno da se funkcija slučajnog uzorka zove ***statistika***, može se reći da je procjenitelj određena statistika.

Formulom (3) definiran je jedan procjenitelj parametra p u 1. primjeru. Jasno je da se mogu definirati i drugi procjenitelji za parametar p, različiti od onog u (3). Može se, naime, uzeti neka druga statistika, recimo $\frac{1}{2}\overline{X} = \frac{1}{2n}(X_1 + \ldots + X_n)$, proglasiti je procjeniteljem nepoznatog parametra p i zatim postaviti zadatak da se međusobno usporede procjenitelji $\overline{X}$ i $\frac{1}{2}\overline{X}$. Može se također postaviti zadatak da se u zadanom skupu procjenitelja pronađe, u određenom smislu, najbolji procjenitelj za parametar p, o čemu će biti riječi kasnije.

Općenito se problem rješava tako da se definira tzv. *funkcija gubitka* (loss function) $(\widehat{t}, t) \mapsto L(\widehat{t}, t)$, čija se vrijednost $L(\widehat{t}, t) \in \mathbf{R}$ može interpretirati kao gubitak ili trošak aproksimiranja nepoznatog parametra t, vrijednošću $\widehat{t}$ procjenitelja $\widehat{T}$, dobivene na temelju niza mjerenja $x_1, \ldots, x_n$. Odmah se vidi da je $L(\widehat{T}, t)$, kao funkcija slučajne varijable $\widehat{T}$, također s.v., pa se može definirati funkcija

(8) $$t \mapsto R(t) = \mathrm{E}[L(\widehat{T}, t)], t \in \Theta,$$

koja se zove *funkcija rizika* (risk function). Broj $R(t)$ označuje očekivani gubitak (trošak, rizik) pri aproksimaciji parametra t vrijednošću $\widehat{t}$.

Graf funkcije $t \mapsto R(t)$ zove se *operativna karakteristika* procjenitelja $\widehat{T}$ za danu funkciju gubitka L.

U 1. primjeru uzeli smo kao funkciju gubitka kvadrat razlike između vrijednosti $\overline{x}$ procjenitelja $\overline{X}$ i nepoznatog parametra p, tj. $L(\overline{x}, p) = (\overline{x} - p)^2$, dok je funkcija rizika izražena formulom (5), a operativna karakteristika prikazana je na sl. 1.

Primijetimo da se u teorijskom modelu pojavljuju s.v., kao što su $X_1, \dots, X_n$, zatim statistika, odnosno procjenitelj $\widehat{T}$, čija vjerojatnosna razdioba nije fiksirana, već je ovisna o parametru $t \in \Theta$. Stoga će i sve izvedene veličine, kao što su na primjer očekivanje, varijanca i sl. za te s.v. također ovisiti o parametru t, pa zato treba voditi računa da se operatori E (očekivanje) i V (varijanca) odnose na vjerojatnosnu razdiobu P_t. Korektnije bi, zapravo, bilo da se oni označuju sa E_t, odnosno V_t, kako bi se istaknula njihova ovisnost o parametru t, ali će se radi jednostavnosti pisanja izostavljati indeks t. Tako smo već postupili u formuli (5), gdje je umjesto precizne oznake $\mathrm{V}_p[\overline{X}]$ upotrijebljena jednostavnija oznaka $\mathrm{V}[\overline{X}]$, a slično ćemo postupati i ubuduće.

Funkcija rizika, odnosno operativna karakteristika, može poslužiti kao određeni pokazatelj pri međusobnom uspoređivanju različitih procjenitelja. Uzme li se u 1. primjeru procjenitelj $a\overline{X}$ $(0 \le a \le 1)$ umjesto procjenitelja $\overline{X}$, i funkcija gubitka $L(a\overline{x}, p) = (a\overline{x} - p)^2$, za funkciju rizika dobiva se

$$R_1(p) = \mathrm{E}[(a\overline{X} - p)^2] = a^2\mathrm{E}[\overline{X}^2] - 2ap\mathrm{E}[\overline{X}] + p^2 .$$

Budući da je $\mathrm{E}[\overline{X}] = p$, $\mathrm{E}[\overline{X}^2] = \mathrm{V}[\overline{X}] + (\mathrm{E}[\overline{X}])^2 = \frac{1}{n}\, p(1-p) + p^2$, dobiva se

$$R_1(p) = \frac{1}{n}\, a^2 p(1-p) + np^2(1-a)^2, p \in \langle 0, 1\rangle . \tag{9}$$

Za $a = 1$ očigledno (9) postaje (5). Usporede li se funkcije rizika (5) i (9), tako da se promotri

$$\frac{R_1(p)}{R(p)} = a^2 + \frac{np(1-a)^2}{1-p},$$

vidi se da jednadžba $a^2 + \frac{np(1-a)^2}{1-p} = 1$ ima rješenje $p = p_0 = \frac{1+a}{n(1-a)+1+a}$, što znači da je $R_1(p_0) = R(p_0)$, dok je za $p < p_0$, $R_1(p) < R(p)$, a za $p > p_0$, $R_1(p) > R(p)$.

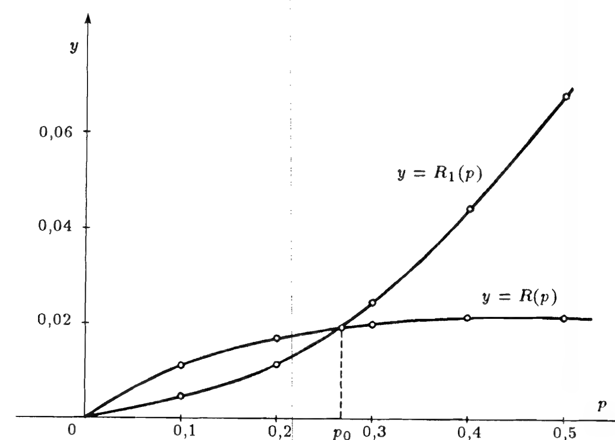
Uzme li se, na primjer, $a = 0{,}5$ dobiva se $p_0 = \frac{3}{n+3}$, što znači da je $\frac{1}{2}\overline{X}$ bolji procjenitelj (u smislu manjeg rizika) od $\overline{X}$, ako je stvarna vrijednost parametra p manja od $\frac{3}{n+3}$. Za velike uzorke $(n \to \infty)$ očigledno je interval $\left\langle 0, \frac{3}{n+3}\right\rangle$, na kojem je $\frac{1}{2}\overline{X}$ bolji procjenitelj od $\overline{X}$, vrlo uzak. No, za male uzorke taj interval nije beznačajan. Za $n = 10$ to je interval $\left\langle 0, \frac{3}{13}\right\rangle$ i tada je

$$R_1(p) = \frac{p}{40}(9p+1),$$

dok je

$$R(p) = \frac{1}{10}(1-p)p.$$

Skica odgovarajućih grafova prikazana je na sl. 2.



Slika 2. Operativne karakteristike procjenitelja $\overline{X}$ i $\frac{1}{2}\overline{X}$ za $n = 10$

Iz (8) se razabire da funkcija rizika, odnosno pripadna operativna karakteristika, osim o izabranom procjenitelju ovisi i o izabranoj funkciji gubitka L. Funkcija gubitka obično se odabire u skladu s prirodom konkretnog problema.

Najčešće se kao funkcija gubitaka pri procjeni nepoznatog parametra t $(t \in \Theta \subseteq \mathbf{R})$ vrijednošću $\widehat{t}$, procjenitelja $\widehat{T}$, uzima kvadratna greška, tj. stavlja se

$$L(\widehat{t}, t) = (\widehat{t} - t)^2, \tag{10}$$

tako da je pripadna funkcija rizika

$$R(t) = \mathrm{E}[(\widehat{T} - t)^2], t \in \Theta. \tag{11}$$

Vrijednost $R(t)$ može se, dakle, interpretirati kao *očekivana* ili *srednja kvadratna greška* pri aproksimaciji nepoznatog parametra t vrijednošću $\widehat{t}$ procjenitelja $\widehat{T}$.

> Ako procjenitelj $\widehat{T}$ zadovoljava uvjet
>
> $$\mathrm{E}[\widehat{T}] = t, \tag{12}$$
>
> onda se kaže da je $\widehat{T}$ ***nepristrani*** ili *centrirani procjenitelj*.

Inače se veličina

$$b(t) = \mathrm{E}[\widehat{T}] - t \tag{13}$$

zove *pristranost* procjenitelja.

Za nepristrani procjenitelj formula (11) postaje

$$R(t) = \mathrm{E}[(\widehat{T} - \mathrm{E}[\widehat{T}])^2] = \mathrm{V}[\widehat{T}], \tag{14}$$

pa se vidi da su vrijednosti funkcije rizika varijance procjenitelja. Prema tome, ako se funkcija rizika definira formulom (11) onda će se, prema formuli (69) iz IV.6, minimalni rizik postići s nepristranim procjeniteljem koji ima minimalnu varijancu.

Općenito se može pisati

$$\mathrm{E}[(\widehat{T}-t)^2] = \mathrm{E}[(\widehat{T}-\mathrm{E}[\widehat{T}]+b(t))^2] = \mathrm{E}[(\widehat{T}-\mathrm{E}[\widehat{T}])^2] + b^2(t),$$

tako da funkcija rizika ima oblik

(15) $$R(t) = \mathrm{V}[\widehat{T}] + b^2(t), t \in \Theta,$$

Procjenitelj $\overline{X}$, razmatran u 1. primjeru i definiran u (3), nepristrani je procjenitelj parametra p, dok procjenitelj $a\overline{X}$ $(0 < a < 1)$ nije nepristran jer je $\mathrm{E}[a\overline{X}] = a\mathrm{E}[\overline{X}] = ap$, pa je njegova pristranost $b(p) = p(a-1)$. Iako je nepristranost poželjno svojstvo procjenitelja, može se dogoditi da pristrani procjenitelj ima, bar u nekom dijelu skupa Θ dopuštenih vrijednosti parametra, manje vrijednosti funkcije rizika nego nepristrani procjenitelj. To se, na primjer, vidi na sl. 2.

Da se pri izboru procjenitelja nije dobro osloniti samo na intuiciju, već da treba imati i egzaktnije metode, pokazat će nam idući primjer.

2. primjer

Vrijeme života pojedinog individuuma u određenoj biološkoj vrsti slučajna je varijabla. Pojedini individuumi umiru već pri rađanju, tako da je najmanje moguće vrijeme života nula vremenskih jedinica. Odmah se može postaviti zadatak da se, na temelju n mjerenja $x_1, \ldots, x_n$ (x_i označuje životni vijek i-tog individuuma), procijeni najveće moguće trajanje života u promatranoj biološkoj vrsti.

Da bi se definirao odgovarajući teorijski model pretpostavit će se da je $(x_1, \ldots, x_n)$ vrijednost slučajnog uzorka $(X_1, \ldots, X_n)$, gdje $X_i \sim \mathrm{U}(0,t)$, $i = 1, \ldots, n$. Drugim riječima, pretpostavlja se da vrijeme života ima uniformnu razdiobu na segmentu $[0,t]$ $(t > 0)$, pa se zadatak svodi na procjenu nepoznatog parametra t u parametarskom modelu s klasom dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{\mathrm{U}(0,t) : 0 < t < \infty\}$.

Pri definiranju procjenitelja za parametar t može se rezonirati, na primjer, ovako: aritmetička sredina $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ trebala bi pasti negdje oko sredine segmenta $[0,t]$, jer je teorijska sredina uniformne razdiobe $\frac{1}{2}t$ (v. IV.6). Stoga bi se nepoznati parametar t mogao aproksimirati vrijednošću $2\overline{x}$, pa se čini da bi

(16) $$\widehat{T}_1 = 2\overline{X} = \frac{2}{n}(X_1 + \ldots + X_n)$$

mogao biti dobar procjenitelj za parametar t. Odmah se vidi (formula (47) u V.6) da je

$$\mathrm{E}[\widehat{T}_1] = \frac{2}{n}\mathrm{E}[X_1 + \ldots + X_n] = \frac{2}{n} n\mathrm{E}[X_1] = 2\frac{t}{2} = t,$$

što znači da je $\widehat{T}_1$ nepristrani procjenitelj. Primjenom formule (49) u V.6. izvodi se odgovarajuća očekivana kvadratna greška

(17) $$R_1(t) = \mathrm{V}[\widehat{T}_1] = \frac{4}{n^2} n\mathrm{V}[X_1] = \frac{4}{n}\frac{t^2}{12} = \frac{1}{3n}t^2, \quad t > 0.$$

Drugi način razmišljanja pokazuje da bi najveća vrijednost u nizu $x_1, \ldots, x_n$ trebala biti blizu nepoznatog parametra t, pa se čini razumnim uzeti slučajnu varijablu

(18) $$\widehat{T}_2 = \max(X_1, \ldots, X_n)$$

kao procjenitelj za nepoznati parametar t.

Primjenom rezultata navedenog u formuli (50), točke 9. u V.6, može se dobiti f.r.v. za s.v. $\widehat{T}_2$. Ovdje, naime, f.r.v. za s.v. $X_i (i = 1, \ldots, n)$ jest

$$F(x) = \begin{cases} 0, & \text{za} \quad x < 0 \\ \dfrac{x}{t}, & \text{za} \quad 0 \le x \le t \\ 1, & \text{za} \quad x > t, \end{cases}$$

iz čega slijedi da f.r.v. za s.v. $\widehat{T}_2$ glasi

(19) $$G(x) = [F(x)]^n = \begin{cases} 0, & \text{za} \quad x < 0 \\ \left(\dfrac{x}{t}\right)^n, & \text{za} \quad 0 \le x \le t \\ 1, & \text{za} \quad x > t, \end{cases}$$

a pripadna f.g.v. glasi

(20) $$g(x) = \frac{\mathrm{d}\,G(x)}{\mathrm{d}\,x} = \begin{cases} 0, & \text{za} \quad x < 0 \text{ i } x > t \\ \dfrac{n}{t^n}x^{n-1}, & \text{za} \quad 0 \le x \le t. \end{cases}$$

To omogućuje da se odredi

(21) $$\begin{cases} \mathrm{E}[\widehat{T}_2] = \displaystyle\int_{-\infty}^{\infty} x g(x)\,\mathrm{d}\,x = \frac{n}{n+1}t, \\ \mathrm{V}[\widehat{T}_2] = \displaystyle\int_{-\infty}^{\infty} x^2 g(x)\,\mathrm{d}\,x - \left(\frac{n}{n+1}t\right)^2 = \frac{t^2 n}{(n+1)^2(n+2)}. \end{cases}$$

Iz (21) se vidi da $\widehat{T}_2$ nije nepristrani procjenitelj i da je njegova pristranost

$$b(t) = -\frac{t}{n+1}.$$

Ako se, primjenom (15), odredi očekivana kvadratna greška (funkcija rizika) za procjenitelj $\widehat{T}_2$, dobiva se

(22) $$R_2(t) = \frac{2t^2}{(n+1)(n+2)}, t > 0.$$

Usporede li se procjenitelji $\widehat{T}_1$ i $\widehat{T}_2$ za nepoznati parametar t, tako da se načini kvocijent

$$\frac{R_2(t)}{R_1(t)} = \frac{6n}{(n+1)(n+2)} = q(n),$$

vidi se da taj kvocijent ne ovisi o t, već samo o n, i to tako da je za $n > 2$ rizik pri procjeni procjeniteljem $\widehat{T}_2$ manji od rizika pri procjeni nepoznatog parametra t procjeniteljem $\widehat{T}_1$.

Tablica 1.

n	1	2	3	5	10	50	100	500
$q(n)$	1	1	0,9	0,71	0,45	0,11	0,06	0,01

Iz tabl. 1. vidi se da je, već za uzorke veličine oko $n = 50$, taj rizik gotovo deset puta manji.

Ovaj primjer donekle objašnjava već ranije izrečenu primjedbu da nepristranost i nije baš bitno svojstvo za dobre procjenitelje. Od pristranog procjenitelja se, inače, jednostavnim "popravkom" može dobiti nepristrani procjenitelj, koji u pogledu rizika ima slična svojstva kao i polazni pristrani procjenitelj.

Uzme li se, na primjer, umjesto pristranog procjenitelja $\widehat{T}_2$ procjenitelj

$$\widehat{T}_3 = \frac{n+1}{n}\widehat{T}_2 = \frac{n+1}{n}\max(X_1, \dots, X_n),$$

dobiva se nepristrani procjenitelj jer je očigledno

$$\mathrm{E}[\widehat{T}_3] = \frac{n+1}{n}\mathrm{E}[\widehat{T}_2] = t.$$

Funkcija rizika za procjenitelj $\widehat{T}_3$ glasi

$$(23) \qquad R_3(t) = \mathrm{V}[\widehat{T}_3] = \frac{n+1}{n}\mathrm{V}[\widehat{T}_2] = \frac{t^2}{n(n+2)}, t > 0,$$

pa se vidi da ona, poput $R_2(t)$, opada kao n^{-2}, za razliku od $R_1(t)$, koji opada kao n^{-1}.

Prethodna razmatranja vrlo očigledno pokazuju kako je složen problem određivanja dobrog procjenitelja. Tako se, općenito, bez dodatnih pretpostavki, i ne može naći procjenitelj koji bi imao uniformno (za svaki $t \in \Theta$) najmanji rizik u skupu svih mogućih procjenitelja. Poznat je, međutim, tzv. *minimaks-princip* za izbor procjenitelja najmanjeg rizika, koji kaže da za procjenu nepoznatog parametra t treba uzeti onaj procjenitelj $\widehat{T}_0$ za koji vrijedi

$$(24) \qquad R_0(t) = \mathrm{E}[(\widehat{T}_0 - t)^2] = \min_{\widehat{T}\in\mathcal{T}}\{\max_{t\in\Theta}\mathrm{E}[(\widehat{T} - t)^2]\},$$

pri čemu je $\mathcal{T}$ određena klasa procjenitelja za parametar t.

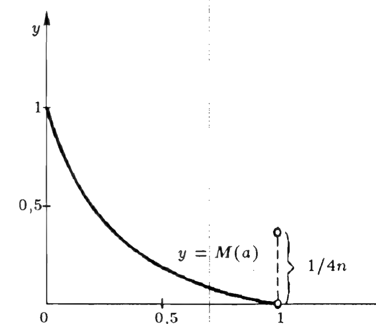
Ako bi se, na primjer, želio naći procjenitelj za parametar p iz 1. primjera u skladu s minimaks-principom, pri čemu se za $\widehat{T}$ dopuštaju procjenitelji oblika $a\overline{X}$, tj. $\mathcal{T} = \{\widehat{T} = a\overline{X} : 0 \le a \le 1\}$, onda bi, prema (9), trebalo promatrati

$$\min_{a\in[0,1]}\left\{\max_{p\in[0,1]}\left[\frac{1}{n}a^2p(1-p) + np^2(1-a)^2\right]\right\}.$$

Budući da je

$$M(a) = \max_{p\in[0,1]}\left[\frac{1}{n}a^2p(1-p) + np^2(1-a)^2\right] = \begin{cases}(1-a)^2, & \text{za } 0 \le a < 1\\ \dfrac{1}{4n}, & \text{za } a = 1,\end{cases}$$

očigledno je da ne postoji $a \in [0,1]$ za koji bi funkcija $a \mapsto M(a), 0 \le a \le 1$,



Slika 3. Skica grafa funkcije $a \mapsto M(a)$

postigla minimalnu vrijednost. Stoga se zaključuje da u klasi procjenitelja oblika $a\overline{X}$ ne postoji procjenitelj za parametar p, koji bi zadovoljio minimaks-princip. Vidjeli smo da $a\overline{X}$, za $a \ne 1$, nije nepristrani procjenitelj za parametar p, pa se može pomisliti da procjenitelj koji zadovoljava minimaks-princip treba tražiti u klasi nepristranih procjenitelja.

Općenito vrijedi da će nepristrani procjenitelj minimalne varijance zadovoljavati minimaks-princip, što se razabire iz formule (69) u IV.6. Takvi procjenitelji obično se zovu *najefikasniji procjenitelji*, a o tome kako se pronalaze bit će riječi u VI.7.

2. Procjena očekivanja i varijance

U mnogim praktičnim situacijama potrebno je procijeniti teorijsku srednju vrijednost (matematičko očekivanje) neke veličine na temelju n ($n \in \mathbf{N}$) izvedenih mjerenja. Poznato je, na primjer, da se ponavljanjem mjerenja težine određenog predmeta redovito ne dobivaju jednaki rezultati, već se dobiveni niz mjerenja $x_1, \dots, x_n$ može razmatrati kao niz statističkih podataka. Stvarna težina μ može se shvatiti kao matematičko očekivanje $\mathrm{E}(X)$ neke s.v. X, pa se problem utvrđivanja stvarne težine dotičnog predmeta može interpretirati i kao problem procjene parametra μ s.v. X, na temelju n nezavisnih mjerenja. Tada se $(x_1, \dots, x_n)$ shvaća

kao vrijednost slučajnog uzorka $(X_1, \ldots, X_n)$, gdje su $X_1, \ldots, X_n$ nezavisne s.v. s istom vjerojatnosnom razdiobom kao i s.v. X.

Teorijski model koji omogućuje rješavanje problema izgrađuje se tako da se, kao klasa dopuštenih razdioba vjerojatnosti za s.v. X, uzme klasa $\mathcal{P}$ koja uključuje sve vjerojatnosne razdiobe s konačnim očekivanjem i fiksiranom varijancom σ^2. Klasa $\mathcal{P}$ ne može se opisati pomoću konačnog broja parametara, tako da je ovdje riječ o neparametarskom modelu.

Pri definiranju procjenitelja za nepoznato očekivanje μ intuicija nas navodi na ideju da bi statistika

$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$

mogla biti dobar procjenitelj za nepoznato očekivanje μ s.v. X. Očigledno je, naime, da bi aritmetička sredina $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$ izmjerenih podataka trebala pasti blizu nepoznatog parametra μ.

Statistika $\overline{X}$ zove se ***uzoračka aritmetička sredina*** i odmah se vidi (formula (47) u V.6) da vrijedi

$$\text{(25)} \qquad \mathrm{E}[\overline{X}] = \frac{1}{n}(\mathrm{E}[X_1] + \ldots + \mathrm{E}[X_n]) = \frac{1}{n}n\mu = \mu,$$

pa se može zaključiti da je $\overline{X}$ nepristrani procjenitelj za μ.
Pripadna srednja kvadratna greška, prema (49) u V.6. glasi

$$\text{(26)} \qquad \mathrm{V}[\overline{X}] = \frac{1}{n^2}(\mathrm{V}[X_1] + \ldots + \mathrm{V}[X_n]) = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2, \mu \in \mathbf{R}.$$

Moglo bi se, dakako, pokušati i s drugim procjeniteljima za nepoznato očekivanje μ. Tako se mogu naći određena opravdanja da se vrijednost $\widehat{m}$ medijana niza statističkih podataka $x_1, \ldots, x_n$ (v. II.2) uzme kao procjena za nepoznati parametar μ i na temelju toga definira procjenitelj $\widehat{M}$ za μ.

Intuitivno bi se moglo opravdati i uzimanje statistike $\widehat{T} = \frac{1}{2}[\min(x_1, \ldots, x_n) + \max(x_1, \ldots, x_n)]$, čija vrijednost $\widehat{t}$ označuje sredinu između najmanje i najveće izmjerene vrijednosti, kao procjenitelja za teorijsku sredinu μ s.v. X.

Međutim, u pretpostavljenome teorijskom modelu praktički je nemoguće utvrditi svojstva ovih i drugih procjenitelja, bitno različitih od procjenitelja $\overline{X}$, na temelju kojih bi se oni mogli međusobno uspoređivati. Za rješavanje takvih problema trebalo bi teorijski model promijeniti, recimo tako da se za s.v. X dopuste samo normalne razdiobe, tj. da se stavi $\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma > 0\}$. No time se izlažemo opasnosti da nam stvarna vjerojatnosna raspodjela ostane izvan pretpostavljenog modela.

U statističkim zaključivanjima uvijek postoji dilema o tome da li teorijski model definirati tako da klasa $\mathcal{P}$ dopuštenih vjerojatnosnih razdioba bude što opsežnija, kako bi bio što manji rizik da stvarna razdioba ne pripada klasi $\mathcal{P}$, ili pak klasu detaljnije opisati, što znači da će biti manje opsežna, ali će omogućiti više teorijskih spoznaja o odnosu nepoznatog parametra i njegovih procjenitelja.

Stvar je kreatora teorijskog modela da, na temelju znanja, iskustva i intuicije, nađe razumnu ravnotežu između tih kontradiktornih zahtjeva.

U teoriji statističkog zaključivanja razrađeni su određeni teorijski modeli i stručnjak se odlučuje za primjenu odgovarajućeg modela u konkretnoj situaciji.

Za problem procjene nepoznate varijance $\mathrm{V}[X] = \sigma^2$, teorijski model sastojat će se od slučajnog uzorka veličine $n > 4$ i klase $\mathcal{P}$ dopuštenih vjerojatnosnih razdioba u koju su uključene sve one razdiobe koje imaju fiksirani konačni četvrti centralni moment μ_4. Opet se čini razumnim definirati procjenitelj za nepoznati parametar σ^2 tako da se σ^2 aproksimira vrijednošću uzoračke varijance $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (v. II.3). To znači da će se statistika

$$\text{(27)} \qquad \widehat{\Sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

koja se inače zove ***uzoračka varijanca***, uzeti kao procjenitelj za nepoznati parametar σ^2.

Budući da je $\widehat{\Sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$, bit će

$$\mathrm{E}[\widehat{\Sigma}^2] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i^2] - \mathrm{E}[\overline{X}^2],$$

što se, na temelju formule (33) u IV.4, može pisati i

$$\mathrm{E}[\widehat{\Sigma}^2] = \frac{1}{n}\sum_{i=1}^{n}\{\mathrm{V}[X_i] + (\mathrm{E}[X_i])^2\} - \mathrm{V}[\overline{X}] - (\mathrm{E}[\overline{X}])^2.$$

Uzevši još u obzir da je $\mathrm{E}[X_i] = \mu, \mathrm{V}[X_i] = \sigma^2 (i = 1, \ldots, n), \mathrm{E}[\overline{X}] = \mu$ i $\mathrm{V}[\overline{X}] = \frac{1}{n}\sigma^2$, konačno se dobiva

$$\text{(28)} \qquad \mathrm{E}[\widehat{\Sigma}^2] = \frac{1}{n}[n(\sigma^2 + \mu^2)] - \frac{1}{n}\sigma^2 - \mu^2 = \frac{n-1}{n}\sigma^2.$$

Iz (28) se vidi da $\widehat{\Sigma}^2$ nije nepristrani procjenitelj za σ^2, ali će statistika

$$\text{(29)} \qquad S^2 = \frac{n}{n-1}\widehat{\Sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

koja se inače zove ***korigirana uzoračka varijanca***, biti nepristrani procjenitelj nepoznate varijance σ^2.

Nešto složenijim izvodom (v. zad. 3) pokazuje se da je

$$\text{(30)} \qquad \mathrm{V}[S^2] = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

dok je

(31) $$V[\hat{\Sigma}^2] = \frac{1}{n}\left[\left(1-\frac{1}{n}\right)^2\mu_4 - \left(1-\frac{1}{n}\right)\left(1-\frac{3}{n}\right)\sigma^4\right].$$

3. Metoda najveće vjerojatnosti

Prethodno razmatranje pokazalo je da definiranje procjenitelja u problemu procjene parametara ima ključnu ulogu. S druge strane, vidjelo se da intuicija baš i nije uvijek pouzdan voditelj pri definiranju procjenitelja, pa se prirodno nameće ideja da se nađu neka opća načela na temelju kojih bi se pronalazili dobri procjenitelji. Jedno od najvažnijih i najplodotvornijih načela zove se ***metoda najveće vjerojatnosti*** (engleski: *Maximum Likelihood Method*), ili kraće ***ML-metoda***. Za bolje razumijevanje biti ove metode razmotrit će se idući primjer.

3. primjer

Uzmimo da je u 1. primjeru $n = 10$ i da dobiveni niz od 10 mjerenja glasi 0,1,0,1,0,0,0,1,0,0. Vjerojatnost da se mjerenjem dobije baš taj niz, uz pretpostavljeni teorijski model, iznosi

$$\begin{aligned}&P(X_1=0,X_2=1,X_3=0,X_4=1,X_5=0,X_6=0,X_7=0,X_8=1,X_9=0,X_{10}=0)=\\&=P(X_1=0)P(X_2=1)P(X_3=0)P(X_4=1)P(X_5=0)P(X_6=0)P(X_7=0)P(X_8=1)\cdot\\&\cdot P(X_9=0)P(X_{10}=0)=p^3(1-p)^7=\boldsymbol{L}(p).\end{aligned}$$

Promotri li se funkcija $p \mapsto \boldsymbol{L}(p), p \in [0,1]$, vidi se da je $\boldsymbol{L}(0) = \boldsymbol{L}(1) = 0$ i da ona poprima maksimalnu vrijednost $\boldsymbol{L}(0{,}3) = 0{,}3^3 \cdot 0{,}7^7 \approx 0{,}0022$, baš za $p = 0{,}3$. (To se može jednostavno izvesti primjenom diferencijalnog računa; v. zad. 8.) Stoga se čini opravdanim vjerovati da je stvarna vrijednost nepoznatog parametra p baš 0,3 a ne recimo 0,1 ili 0,6, jer je $\boldsymbol{L}(0{,}1) \approx 0{,}0005$, dok je $\boldsymbol{L}(0{,}6) \approx 0{,}00035$, što su znatno manje vrijednosti od $\boldsymbol{L}(0{,}3)$.

Kada bi, dakle, stvarna vrijednost nepoznatog parametra bila $p = 0{,}3$, onda teorija pokazuje da je vjerojatnost dobivanja baš izmjerenog niza podataka najveća. To nas upućuje da vrijednost 0,3 uzmemo kao procjenu za nepoznati parametar p.

Opće načelo može se formulirati ovako: Da bi se u parametarskom modelu s klasom dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{P_t : t \in \Theta\}$ definirao procjenitelj za nepoznati parametar t, na temelju niza podataka $x_1, \ldots, x_n$, konstruirat će se funkcija

(32) $$t \mapsto \boldsymbol{L}(t) = P_t(X_1 = x_1) \cdot \ldots \cdot P_t(X_n = x_n), t \in \Theta,$$

i odrediti (ako postoji) ona vrijednost $t = \hat{t} \in \Theta$ za koju ta funkcija poprima najveću vrijednost i $\hat{t}$ će se uzeti kao procjena za nepoznati parametar t.

Očigledno je da $\boldsymbol{L}(t)$ ima značenje vjerojatnosti da se pri mjerenju slučajne varijable X dobije baš dani niz podataka, ako je stvarna vrijednost parametra baš t. Stoga se to načelo i zove metoda najveće vjerojatnosti, jer se kao procjena

nepoznatog parametra uzima ona vrijednost t iz skupa Θ svih mogućih vrijednosti koja pojavljivanje baš izmjerenog niza podataka čini najvjerojatnijim.

U 3. primjeru imali smo $t = p$ i riječ je bila o klasi $\mathcal{P} = \{B(1,p) : p \in \langle 0,1\rangle\}$ Bernoullijevih razdioba nepoznatog parametra p, koje pripadaju u diskretne razdiobe vjerojatnosti (v. IV.1. i IV.2), što je omogućilo da se dobije eksplicitni izraz za funkciju $p \mapsto \boldsymbol{L}(p)$. Slično će se postupiti i za bilo koju drugu klasu diskretnih vjerojatnosnih razdioba.

Ako je, međutim, riječ o klasi $\mathcal{P}$ kontinuiranih razdioba vjerojatnosti, onda je $P_t(X_i = x_i) = 0, i = 1, \ldots, n$, pa se iz (32) vidi da je tada $\boldsymbol{L}(t) = 0$ za svaki $t \in \Theta$ pa opisano načelo, očigledno, u tom obliku ne funkcionira.

Statistička zakonitost u kontinuiranoj razdiobi vjerojatnosti, kao što je već rečeno, izražava se pomoću f.g.v., pa se prirodno nameće ideja da se u tom slučaju $\boldsymbol{L}(t)$ definira formulom

(33) $$\boldsymbol{L}(t) = f_t(x_1) \cdot \ldots \cdot f_t(x_n), t \in \Theta,$$

gdje f_t označuje f.g.v. kontinuirane vjerojatnosne razdiobe P_t. Stoga se $\boldsymbol{L}(t)$, definirano u (33), može interpretirati kao gustoća vjerojatnosti slučajnog vektora $(X_1, \ldots, X_n)$ u točki $(x_1, \ldots, x_n) \in \mathbf{R}^n$. ML-metoda i sada funkcionira tako da se odredi (ako postoji) ono $t = \hat{t}$, za koje funkcija (33) postiže najveću vrijednost.

Za procjenu nepoznatog parametra t kontinuirane vjerojatnosne razdiobe uzima se, dakle, ona vrijednost $\hat{t} \in \Theta$ koja gustoći vjerojatnosti u točki $(x_1, \ldots, x_n) \in \mathbf{R}^n$ daje najveću vrijednost.

Općenito se funkcija $t \mapsto \boldsymbol{L}(t)$ definirana u (32), odnosno (33), zove *funkcija vjerodostojnosti* (engleski: *likelihood function*).

4. primjer

U 2. primjeru riječ je bila o procjeni parametra t ($t > 0$) uniformne razdiobe $U(0,t)$, pa se može postaviti zadatak da se nađe odgovarajući procjenitelj ML-metodom. Ovdje pripadna f.g.v. jest

$$f_t(x) = \begin{cases} \dfrac{1}{t}, & \text{za} \quad 0 \le x \le t \\ 0, & \text{za} \quad x < 0 \text{ i } x > t, \end{cases}$$

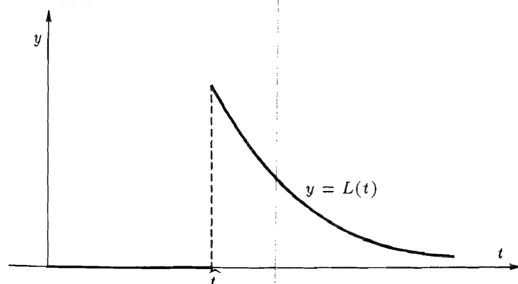
pa je

(34) $$f_t(x_1) \cdot \ldots \cdot f_t(x_n) = \begin{cases} \dfrac{1}{t^n}, & \text{za} \quad 0 \le \max(x_1, \ldots, x_n) \le t \\ 0, & \text{za ostale} \quad (x_1, \ldots, x_n) \in \mathbf{R}^n. \end{cases}$$

Iz (33) i (34) odmah slijedi da odgovarajuća funkcija vjerodostojnosti glasi

(35) $$\boldsymbol{L}(t) = \begin{cases} \dfrac{1}{t^n}, & \text{za} \quad t \ge \hat{t} \\ 0, & \text{za} \quad 0 < t < \hat{t}, \end{cases}$$

gdje je $\hat{t} = \max(x_1, \ldots, x_n)$.



Slika 4. Skica grafa funkcije definirane u (35)

Iz sl. 4. zorno se vidi da funkcija vjerodostojnosti definirana u (35) postiže najveću vrijednost za $t = \widehat{t} = \max(x_1, \ldots, x_n)$, što znači da je $\widehat{t}$ procjena za nepoznati parametar t u smislu ML-metode. (Primijetimo da se taj rezultat ne može izvesti primjenom tehnike diferencijalnog računa.)

U 2. primjeru razmotrena su svojstva procjenitelja $\widehat{T}_2$, definiranog formulom (18), pa se sada može primijetiti da je $\widehat{t}$, zapravo, vrijednost procjenitelja $\widehat{T}_2$. Stoga će se procjenitelj $\widehat{T}_2$ zvati ML-procjenitelj nepoznatog parametra t uniformne razdiobe $U(0, t)$.

Razmatranja u 2. primjeru pokazala su da ML-procjenitelj $\widehat{T}_2$ ima mnogo bolja svojstva od procjenitelja $\widehat{T}_1$ (definiranog formulom (16)) istog parametra t. Kasnije će se pokazati da ML-procjenitelji općenito imaju neka dobra svojstva, koja im daju prednost pred procjeniteljima dobivenima na temelju nekih drugih načela.

Određivanje ML-procjenitelja u nekom modelu s klasom dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{P_{\boldsymbol{t}} : \boldsymbol{t} \in \boldsymbol{\Theta}\}$, gdje $\boldsymbol{t}$ može biti i vektorski parametar, tj. $\boldsymbol{t} = (t_1, \ldots, t_k) \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ $(k \in \mathbb{N})$, obično funkcionira na ovaj način: Ako je riječ o klasi diskretnih razdioba vjerojatnosti, onda se funkcija vjerodostojnosti definira formulom

(36a) $$\boldsymbol{L}(\boldsymbol{t}) = P_{\boldsymbol{t}}(X_1 = x_1) \cdot \ldots \cdot P_{\boldsymbol{t}}(X_n = x_n), \quad \boldsymbol{t} \in \boldsymbol{\Theta},$$

a ako je riječ o klasi kontinuiranih razdioba, definira se formulom

(36b) $$\boldsymbol{L}(\boldsymbol{t}) = f_{\boldsymbol{t}}(x_1) \cdot \ldots \cdot f_{\boldsymbol{t}}(x_n), \quad \boldsymbol{t} \in \boldsymbol{\Theta}.$$

Procjenitelj $\widehat{\boldsymbol{T}}$, čije se vrijednosti $\widehat{\boldsymbol{t}}$ dobivaju kao one vrijednosti za koje funkcija vjerodostojnosti postiže maksimum, zove se ***ML-procjenitelj*** nepoznatog parametra $\boldsymbol{t}$.

U mnogim važnim primjerima $\boldsymbol{t} \mapsto \boldsymbol{L}(\boldsymbol{t})$ je diferencijabilna funkcija, koja najveću vrijednost poprima u točki $\widehat{\boldsymbol{t}} \in \mathbb{R}^k$, za koju vrijedi

(37) $$\frac{\partial \boldsymbol{L}(\widehat{\boldsymbol{t}})}{\partial t_1} = \ldots = \frac{\partial \boldsymbol{L}(\widehat{\boldsymbol{t}})}{\partial t_k} = 0.$$

Rješavanjem sustava jednadžbi (37) dobiva se rješenje $\widehat{\boldsymbol{t}} = (\widehat{t}_1, \ldots, \widehat{t}_k)$ izraženo u ovisnosti o izmjerenim podacima $(x_1, \ldots, x_n)$, tako da se može pisati

(38) $$\begin{aligned} \widehat{t}_1 &= h_1(x_1, \ldots, x_n) \\ &\vdots \\ \widehat{t}_k &= h_k(x_1, \ldots, x_n). \end{aligned}$$

Ako se u (38) konkretna vrijednost $(x_1, \ldots, x_n)$ zamijeni slučajnim vektorom $(X_1, \ldots, X_n)$, onda se za statistiku

$$\widehat{T}_j = h_j(X_1, \ldots, X_n), j = 1, \ldots, k,$$

kaže da je *ML-procjenitelj nepoznatog parametra* t_j.

5. primjer

Treba naći opći oblik ML-procjenitelja za parametar λ $(\lambda > 0)$ Poissonove razdiobe $\mathrm{Po}(\lambda)$ (v. IV.3), uz pretpostavku da se raspolaže s nizom podataka $x_1, \ldots, x_n$. Klasa dopuštenih razdioba vjerojatnosti je, dakle, $\mathcal{P} = \{\mathrm{Po}(\lambda) : \lambda > 0\}$, pa pripadna funkcija vjerodostojnosti glasi

$$\begin{aligned} \boldsymbol{L}(\lambda) &= \frac{\lambda^{x_1}}{x_1!} \exp(-\lambda) \ldots \frac{\lambda^{x_n}}{x_n!} \exp(-\lambda) = \\ &= \frac{1}{x_1! \cdot \ldots \cdot x_n!} \lambda^{x_1 + \ldots + x_n} \exp(-n\lambda), \ \lambda > 0. \end{aligned}$$

Stavi li se $\frac{1}{x_1! \cdot \ldots \cdot x_n!} = K$ i $x_1 + \ldots + x_n = n\overline{x}$, dobiva se

$$\boldsymbol{L}(\lambda) = K\lambda^{n\overline{x}} \exp(-n\lambda), \lambda > 0.$$

Deriviranjem po λ dobiva se

$$\frac{\mathrm{d}\,\boldsymbol{L}(\lambda)}{\mathrm{d}\,\lambda} = Kn\lambda^{n\overline{x}-1} \exp(-n\lambda)(\overline{x} - \lambda).$$

Budući da je $Kn\lambda^{n\overline{x}-1} \exp(-n\lambda) > 0$, ostaje da se po λ riješi jednadžba $\overline{x} - \lambda = 0$, iz čega proizlazi $\lambda = \widehat{\lambda} = \overline{x}$.

Prema tome, opći je oblik ML-procjenitelja za nepoznati parametar λ Poissonove razdiobe

(39) $$\widehat{\Lambda} = \overline{X} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Lako se pokazuje da je $\widehat{\Lambda}$ nepristrani procjenitelj i da pripadna funkcija rizika glasi

(40) $$R(\lambda) = \frac{1}{n}\lambda, \quad \lambda > 0.$$

6. primjer

Često se usvaja pretpostavka da je vijek trajanja nekoga tehničkog uređaja slučajna varijabla eksponencijalne razdiobe (v. IV.5). Mjerenjem vijeka trajanja n takvih uređaja dobiven je niz podataka $x_1, \ldots, x_n$. Treba procijeniti nepoznati parametar α ($\alpha > 0$) eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$. Pokušajmo, dakle, u jednoparametarskom modelu s klasom dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{\mathrm{Ex}(\alpha) : \alpha > 0\}$ pronaći ML-procjenitelj za nepoznati parametar α.

Budući da f.g.v. za $\mathrm{Ex}(\alpha)$ glasi

$$f_\alpha(x) = \begin{cases} 0, & \text{za } x \leq 0 \\ \alpha \exp(-\alpha x), & \text{za } x > 0, \end{cases}$$

pripadna funkcija vjerodostojnosti glasi

$$\boldsymbol{L}(\alpha) = \alpha^n \exp[-\alpha(x_1 + \ldots + x_n)], \alpha > 0.$$

Stavi li se $x_1 + \ldots + x_n = n\overline{x}$, dobiva se

$$\boldsymbol{L}(\alpha) = \alpha^n \exp(-\alpha n \overline{x}), \alpha > 0,$$

iz čega, deriviranjem po α, odmah proizlazi

$$\frac{\mathrm{d}\,\boldsymbol{L}(\alpha)}{\mathrm{d}\,\alpha} = n\alpha^{n-1} \exp(-\alpha n \overline{x})(1 - \alpha \overline{x}).$$

Budući da je $n\alpha^{n-1} \exp(-\alpha n\overline{x}) > 0$, ostaje da se po α riješi jednadžba $1 - \alpha\overline{x} = 0$, iz čega se dobiva

(41) $$\alpha = \widehat{\alpha} = \frac{1}{\overline{x}}.$$

Prema tome, opći oblik ML-procjenitelja za nepoznati parametar α eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$ glasi

(42) $$\widehat{A} = \frac{1}{\overline{X}} = \frac{n}{X_1 + \ldots + X_n}.$$

Stvar se može pojednostavniti tako da se $\frac{1}{\alpha} = \alpha_0$ razmatra kao parametar eksponencijalne razdiobe i tada je $\widehat{\alpha}_0 = \overline{x}$, pa opći oblik ML-procjenitelja za parametar α_0 glasi

(43) $$\widehat{A}_0 = \overline{X} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Budući da je $\mathrm{E}[\widehat{A}_0] = \mathrm{E}[\overline{X}] = \frac{1}{n} n \mathrm{E}[X_1] = \mathrm{E}[X_1] = \frac{1}{\alpha} = \alpha_0$, onda je $\widehat{A}_0$ nepristrani procjenitelj za parametar α_0 i pripadna funkcija rizika izgleda

(44) $$R(\alpha_0) = \mathrm{V}[\widehat{A}_0] = \frac{1}{n^2} n \mathrm{V}[X_1] = \frac{1}{n}\alpha_0^2, \quad \alpha_0 > 0.$$

Da je, umjesto jednoparametarskog modela s klasom eksponencijalnih razdioba, usvojen dvoparametarski model s klasom $\mathcal{P} = \{\mathrm{G}(\alpha, \beta) : \alpha > 0, \beta > 0\}$ gama-razdioba s parametrima α i β (v. IV.5), došlo bi se do funkcije vjerodostojnosti oblika

$$\boldsymbol{L}(\alpha, \beta) = \left[\frac{\alpha^\beta}{\Gamma(\beta)}\right]^n (x_1 \cdot \ldots \cdot x_n)^{\beta - 1} \exp[-\alpha(x_1 \cdot \ldots \cdot x_n)].$$

Pokuša li se riješiti sustav jednadžbi

$$\frac{\partial \boldsymbol{L}(\alpha, \beta)}{\partial \alpha} = 0, \frac{\partial \boldsymbol{L}(\alpha, \beta)}{\partial \beta} = 0,$$

po α i β, nailazi se na velike teškoće, jer se α i β ne mogu eksplicite izraziti u ovisnosti o $x_1, \ldots, x_n$, tako da u ovom slučaju ML-metoda ne omogućuje dobivanje općeg izraza za procjenitelje nepoznatih parametara α i β gama-razdiobe $\mathrm{G}(\alpha, \beta)$. To nas upućuje da valja razmotriti i druge metode za dobivanje procjenitelja, o čemu će biti riječi kasnije.

4. Procjenitelji parametara normalne razdiobe

Normalna ili Gaussova razdioba (v. IV.5) ima istaknuto mjesto u matematičkoj statistici i stoga će se detaljnije iznijeti problem procjene parametara μ i σ^2 normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$. Polazi se, dakle, od pretpostavke da je dan niz podataka (mjerenja s.v. X) $x_1, \ldots, x_n$ i da je klasa dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma > 0\}$, tj. da je riječ o dvoparametarskom modelu u kojem je $\boldsymbol{\Theta} = \{(\mu, \sigma^2) \in \mathbf{R}^2 : \mu \in \mathbf{R}, \sigma > 0\}$ skup dopuštenih vrijednosti vektorskog parametra $\boldsymbol{t} = (\mu, \sigma^2)$.

Iz (36b) proizlazi da pripadna funkcija vjerodostojnosti glasi

(45) $$\begin{aligned} \boldsymbol{L}(\boldsymbol{t}) = \boldsymbol{L}(\mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_1 - \mu}{\sigma}\right)^2\right] \cdot \ldots \\ &\ldots \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_n - \mu}{\sigma}\right)^2\right] = \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right], \quad \boldsymbol{t} \in \boldsymbol{\Theta}. \end{aligned}$$

Jednadžba $\frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \mu} = 0$, nakon sređivanja, postaje

$$\sum_{i=1}^n (x_i - \mu) = 0,$$

a jednadžba $\dfrac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \sigma^2} = 0$ postaje

$$-n\sigma^2 + \sum_{i=1}^{n}(x_i - \mu)^2 = 0,$$

što, kao rješenje za μ, daje

$$\mu = \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$

a kao rješenje za σ^2 dobiva se

$$\sigma^2 = \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Stoga su ML-procjenitelji za nepoznate parametre μ i σ^2 normalne razdiobe $N(\mu, \sigma^2)$ statistike

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{i} \quad \widehat{\Sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Može se, dakle, reći da je procjenitelj za μ uzoračka aritmetička sredina $\overline{X}$, a procjenitelj za σ^2 uzoračka varijanca $\widehat{\Sigma}^2$.

Na temelju (25) i (28) vidi se da je $\overline{X}$ nepristrani procjenitelj za parametar μ, dok $\widehat{\Sigma}^2$ nije nepristrani procjenitelj za parametar σ^2, ali se može, kao i u (29), definirati odgovarajući nepristrani procjenitelj

$$S^2 = \frac{n}{n-1}\widehat{\Sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

koji smo nazvali korigirana uzoračka varijanca.

Da bi se dobila funkcija rizika mora se imati na umu da je $\boldsymbol{t} = (\mu, \sigma^2)$ i da će se funkcija gubitka opet definirati kao srednja kvadratna greška pri aproksimaciji vektorskog parametra $\boldsymbol{t} = (\mu, \sigma^2)$ vrijednošću $\widehat{\boldsymbol{t}} = (\widehat{\mu}, \widehat{\sigma}^2)$, tj. kao kvadrat udaljenosti točaka $\boldsymbol{t}$ i $\widehat{\boldsymbol{t}}$ u prostoru $\mathbf{R}^2$. Dobiva se

(46) $$R(\boldsymbol{t}) = R(\mu, \sigma^2) = E[(\overline{X} - \mu)^2] + E[(\widehat{\Sigma}^2 - \sigma^2)^2], \boldsymbol{t} \in \boldsymbol{\Theta}.$$

Iz (26) se razabire da je

(47) $$E[(\overline{X} - \mu)^2] = V[\overline{X}] = \frac{1}{n}\sigma^2,$$

dok iz (28),(31) i činjenice da je u normalnoj razdiobi $N(\mu, \sigma^2)$ četvrti centralni moment $\mu_4 = 3\sigma^4$, proizlazi da je

(48) $$E[(\widehat{\Sigma}^2 - \sigma^2)^2] = \frac{2n-1}{n^2}\sigma^4.$$

Iz (46),(47) i (48) naposljetku se dobiva

(49) $$R(\boldsymbol{t}) = R(\mu, \sigma^2) = \frac{1}{n}\sigma^2 + \frac{2n-1}{n^2}\sigma^4 \ , \ \boldsymbol{t} \in \boldsymbol{\Theta},$$

iz čega se razabire da funkcija rizika, pri procjeni nepoznatoga vektorskog parametra $\boldsymbol{t} = (\mu, \sigma^2)$ vektorskim procjeniteljem $\widehat{\boldsymbol{T}} = (\overline{X}, \widehat{\Sigma}^2)$, ne ovisi o μ, već samo o σ^2. To je i za očekivati s obzirom na značenje parametra μ kao određenog pokazatelja lokacije vjerojatnosne razdiobe, dok je "slučajnost" razdiobe izražena parametrom σ^2.

Ako se $\widehat{\boldsymbol{T}}_1 = (\overline{X}, S^2)$ uzme kao procjenitelj za nepoznati parametar $\boldsymbol{t} = (\mu, \sigma^2)$, onda će odgovarajuća funkcija rizika glasiti

$$R_1(\boldsymbol{t}) = E[(\overline{X} - \mu)^2] + E[(S^2 - \sigma^2)^2], \boldsymbol{t} \in \boldsymbol{\Theta}.$$

Vodeći računa o činjenici da je S^2 nepristrani procjenitelj za σ^2 i da je $E[(S^2 - \sigma^2)^2] = V[S^2]$, te uzimajući u obzir da je $\mu_4 = 3\sigma^4$, iz (30) se dobiva

(50) $$E[(S^2 - \sigma^2)^2] = \frac{2}{n-1}\sigma^4,$$

što zajedno sa (47) naposljetku daje

(51) $$R_1(\boldsymbol{t}) = R_1(\mu, \sigma^2) = \frac{1}{n}\sigma^2 + \frac{2}{n-1}\sigma^4, \boldsymbol{t} \in \boldsymbol{\Theta}.$$

Usporede li se funkcije rizika (49) i (51), vidi se da je, za $n > 1$, srednja kvadratna greška pri aproksimaciji nepoznatih parametara μ i σ^2 vrijednostima ML-procjenitelja $\overline{X}$ i $\widehat{\Sigma}^2$ nešto manja nego pri aproksimaciji vrijednostima nepristranih procjenitelja $\overline{X}$ i S^2. Međutim, za velike uzorke ($n \to \infty$) ta prednost ML-procjenitelja iščezava.

Procjenitelji $\overline{X}$, $\widehat{\Sigma}^2$ i S^2 imaju zanimljiva svojstva i kao određene slučajne varijable, dobivene kao funkcije slučajnog vektora $(X_1, \ldots, X_n)$ u kojem su komponente nezavisne s.v. sa zajedničkom vjerojatnosnom razdiobom $N(\mu, \sigma^2)$. Na temelju točke 1. u V.6. odmah, naime, proizlazi da vrijedi

(52) $$\overline{X} \sim N(\mu, \frac{1}{n}\sigma^2).$$

Teže je dokazati (v. [38]) da vrijedi i

(53) $$\frac{n}{\sigma^2}\widehat{\Sigma}^2 = \frac{n-1}{\sigma^2}S^2 = U \sim \chi^2(n-1),$$

a dokazuje se i da su $\overline{X}$ i U nezavisne slučajne varijable.

Metodom najveće vjerojatnosti može se doći i do procjenitelja za parametre dvodimenzionalne normalne razdiobe $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. Polazi se od pretpostavke da je dan niz dvodimenzionalnih podataka $(x_1, y_1), \ldots, (x_n, y_n)$ i da je klasa dopuštenih razdioba vjerojatnosti

$$\mathcal{P} = \{\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) : (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) \in \boldsymbol{\Theta}\},$$

pri čemu je

$$\boldsymbol{\Theta} = \{\boldsymbol{t} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) \in \mathbf{R}^5 : \mu_1, \mu_2 \in \mathbf{R}, \sigma_1 > 0, \sigma_2 > 0, 0 \leq |\rho| < 1\}.$$

U ovome teorijskom modelu slučajni uzorak veličine n je niz nezavisnih dvodimenzionalnih slučajnih vektora $(X_1, Y_1), \ldots, (X_n, Y_n)$, pri čemu $(X_i, Y_i) \sim \sim \mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, $(i = 1, \ldots, n)$.

Na temelju formula (18), (19) i (20) u V.3. razabire se da pripadna funkcija vjerodostojnosti glasi

$$\boldsymbol{L}(\boldsymbol{t}) = \boldsymbol{L}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = K^n \exp[-\sum_{i=1}^{n} Q(x_i, y_i)].$$

Rješavanjem sustava jednadžbi

$$\frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \mu_1} = 0, \quad \frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \mu_2} = 0, \quad \frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \sigma_1^2} = 0, \quad \frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \sigma_2^2} = 0, \quad \frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \rho} = 0,$$

dolazi se do rješenja

$$\mu_1 = \widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$

$$\mu_2 = \widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y},$$

$$\sigma_1^2 = \widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 = s_x^2,$$

$$\sigma_2^2 = \widehat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 = s_y^2,$$

$$\rho = \widehat{\rho} = \frac{1}{n\widehat{\sigma}_1\widehat{\sigma}_2}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = r.$$

Dobiveni rezultat intuitivno je vrlo prihvatljiv, jer se kao procjene dobivaju upravo one veličine koje su navedene u III.3. i III.4. kao odgovarajući parametri za opisivanje niza statističkih podataka o dvodimenzionalnome statističkom obilježju.

Prema tome, $\widehat{\boldsymbol{T}} = (\overline{X}, \overline{Y}, \widehat{\Sigma}_1^2, \widehat{\Sigma}_2^2, \widehat{P})$ je vektorski ML-procjenitelj za nepoznati vektorski parametar $\boldsymbol{t} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ dvodimenzionalne normalne razdiobe $\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, pri čemu je

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{54}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{55}$$

$$\widehat{\Sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2, \tag{56}$$

$$\widehat{\Sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \overline{Y})^2, \tag{57}$$

$$\widehat{P} = \frac{1}{n\widehat{\Sigma}_1\widehat{\Sigma}_2}\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}). \tag{58}$$

Općenito se statistika $\widehat{P}$, definirana u (58) zove ***uzorački koeficijent korelacije***.

Statistike $\overline{X}$ i $\overline{Y}$, kao procjenitelji za μ_1 i μ_2 imaju svojstva koja su već navedena u (47) i (52), dok statistike $\widehat{\Sigma}_1^2$ i $\widehat{\Sigma}_2^2$, kao procjenitelji za σ_1^2 i σ_2^2 imaju svojstva navedena u (48) i (53). Svojstva procjenitelja $\widehat{P}$ mnogo su složenija i neke jednostavnije formule mogu se izvesti tek uz pretpostavku da je veličina uzorka dovoljno velika ($n \to \infty$), o čemu će biti riječi kasnije (v. VI.8).

5. Metoda momenata

U 6. primjeru vidjelo se da metoda najveće vjerojatnosti dobro ne funkcionira pri procjeni parametara gama-razdiobe $\mathrm{G}(\alpha, \beta)$, pa se stoga prirodno nameće ideja da se za taj slučaj i slične slučajeve pokuša pronaći neko drugo načelo koje bi, dakako, trebalo biti racionalno utemeljeno, a koje bi omogućilo dobivanje jednostavnog rješenja. Jedno takvo načelo realizira se *metodom momenata*, koja se temelji na uvjerenju da su vrijednosti uzoračkih momenata, tj. vrijednosti statističkih momenata (v. II.6) izračunanih na danom nizu podataka $x_1, \ldots, x_n$, bliske vrijednostima teorijskih momenata (v. IV.2. i IV.4) vjerojatnosne razdiobe P_t koja je pretpostavljena u teorijskom modelu. To omogućuje da se formira sustav jednadžbi u kojima je na jednoj strani izraz za dotični teorijski moment, a na drugoj strani vrijednost odgovarajućega statističkog (uzoračkog) momenta.

Tako, na primjer, teorijski izraz za prvi ishodišni moment $\beta_1 = \mathrm{E}[X]$ eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$ je $\frac{1}{\alpha}$, dok je odgovarajući statistički moment $b_1 = \overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$. Iz jednadžbe $\frac{1}{\alpha} = \overline{x}$ proizlazi $\alpha = \frac{1}{\overline{x}}$, što upućuje na to da se statistika $\widehat{A} = \frac{1}{\overline{X}}$ uzme kao procjenitelj za nepoznati parametar α eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$ u smislu metode momenata.

Zanimljivo je primijetiti da se isti rezultat dobio i metodom najveće vjerojatnosti (v. 6. primjer).

Pri procjeni nepoznatih parametara α i β gama-razdiobe $\mathrm{G}(\alpha, \beta)$ iskoristit će se teorijski izraz za prvi ishodišni moment $\beta_1 = \mathrm{E}[X] = \frac{\beta}{\alpha}$ i teorijski izraz za

drugi centralni moment $\mu_2 = D[X] = \frac{\beta}{\alpha^2}$ gama-razdiobe (v. IV.5), te odgovarajući statistički momenti $b_1 = \overline{x}$ i $m_2 = \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Iz sustava jednadžbi

$$\frac{\beta}{\alpha} = \overline{x} \quad , \quad \frac{\beta}{\alpha^2} = \widehat{\sigma}^2$$

odmah proizlazi

$$\alpha = \frac{\overline{x}}{\widehat{\sigma}^2} \quad , \quad \beta = \frac{\overline{x}^2}{\widehat{\sigma}^2}.$$

To pokazuje da će statistike

(59)
$$\widehat{A} = \frac{\overline{X}}{\widehat{\Sigma}^2} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

(60)
$$\widehat{B} = \frac{\overline{X}^2}{\widehat{\Sigma}^2} = \frac{1}{n}\frac{\left(\sum_{i=1}^{n} X_i\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

biti procjenitelji nepoznatih parametara α i β gama-razdiobe $G(\alpha, \beta)$ u smislu metode momenata.

Općenito se može reći da metoda momenata za procjenu parametara u teorijskom modelu s klasom dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{P_{\boldsymbol{t}} : \boldsymbol{t} \in \boldsymbol{\Theta}\}$, gdje je $\boldsymbol{t} = (t_1, \dots, t_k) \in \boldsymbol{\Theta} \subseteq \mathbf{R}^k$ $(k \in \mathbf{N})$, funkcionira ovako: Teorijski izraz za moment r-tog reda ν_r vjerojatnosne razdiobe $P_{\boldsymbol{t}}$ određena je funkcija parametra $\boldsymbol{t}$, pa se može pisati $\nu_r = \nu_r(t_1, \dots, t_k)$. Označi li se sa $\widehat{\nu}_r = \widehat{\nu}_r(x_1, \dots, x_n)$ vrijednost odgovarajućega uzoračkog (statističkog) momenta na nizu podataka (mjerenja s.v. X) $x_1, \dots, x_n$, formirat će se sustav od k jednadžbi s nepoznanicama $t_1, \dots, t_k$, koji glasi

(61)
$$\nu_r(t_1, \dots, t_k) = \widehat{\nu}_r(x_1, \dots, x_n), \quad r = 1, \dots, k.$$

Rješenja (ako postoje) toga sustava

(62)
$$t_j = \widehat{t}_j = \widehat{t}_j(x_1, \dots, x_n), \quad j = 1, \dots, k,$$

omogućuju da se definiraju statistike $\widehat{T}_j = \widehat{T}_j(X_1, \dots, X_n), j = 1, \dots, k$, koje se smatraju procjeniteljima za nepoznate parametre t_j u smislu metode momenata. Također se može reći da je $\widehat{\boldsymbol{T}} = (\widehat{T}_1, \dots, \widehat{T}_k)$ vektorski procjenitelj vektorskog parametra $\boldsymbol{t} = (t_1, \dots, t_k)$ u smislu metode momenata.

Tako će se, na primjer, pri procjeni nepoznatih parametara μ i σ^2 normalne razdiobe $N(\mu, \sigma^2)$ metodom momenata uzeti u obzir činjenica da je ishodišni moment prvog reda $\beta_1 = \mu$ i centralni moment drugog reda $\mu_2 = \sigma^2$, pa će odgovarajući sustav jednadžbi glasiti

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \widehat{\sigma}^2,$$

Sustav je već zapisan tako da se razabiru rješenja po nepoznatim parametarima μ i σ^2. Stoga su statistike $\overline{X}$ i $\widehat{\Sigma}^2$ procjenitelji za μ i σ^2 u smislu metode momenata. Primijetimo da su te iste statistike dobivene u VI.4. i kao ML-procjenitelji za parametre μ i σ^2 normalne razdiobe $N(\mu, \sigma^2)$.

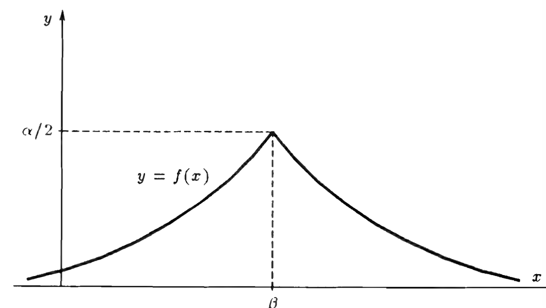
Da metoda momenata i metoda najveće vjerojatnosti mogu dati i posve različite procjenitelje za isti nepoznati parametar pokazat će idući primjer.

7. primjer

U točki 6. iz V.6. definirana je Laplaceova razdioba parametra α, koja se translacijom $h(x) = x + \beta$ prevodi u kontinuiranu vjerojatnosnu razdiobu čija f.g.v. glasi

(63)
$$f(x) = \frac{\alpha}{2}\exp(-\alpha|x - \beta|), x \in \mathbf{R},$$

i koja se zove Laplaceova razdioba s *parametrom oblika* α $(\alpha > 0)$ i *parametrom lokacije* β $(\beta \in \mathbf{R})$.



Slika 5. Skica krivulje Laplaceove razdiobe parametara α i β

Jednostavno se pokazuje da je njezino očekivanje $E[X] = \beta$ i varijanca $V[X] = \frac{2}{\alpha^2}$ (v. zad. 18).

Za procjenu nepoznatih parametara α i β, na temelju niza $x_1, \dots, x_n$ nezavisnih mjerenja s.v. X, formiraju se jednadžbe

$$\beta_1 = E[X] = \beta = \overline{x},$$
$$\mu_2 = V[X] = \frac{2}{\alpha^2} = \widehat{\sigma}^2,$$

čije rješenje po α i β glasi

$$\alpha = \frac{\sqrt{2}}{\hat{\sigma}} \quad , \quad \beta = \overline{x},$$

tako da su statistike

(64) $$\hat{A} = \frac{1}{\hat{\Sigma}}\sqrt{2} \quad , \quad \hat{B} = \overline{X}$$

procjenitelji za parametre α i β Laplaceove razdiobe u smislu metode momenata.

Da bi se odredili ML-procjenitelji za parametre α i β Laplaceove razdiobe treba formirati odgovarajuću funkciju vjerodostojnosti. Iz (36b) i (63) proizlazi

(65) $$\boldsymbol{L}(t) = \boldsymbol{L}(\alpha,\beta) = \left(\frac{\alpha}{2}\right)^n \exp\left(-\alpha \sum_{i=1}^{n} |x_i - \beta|\right), \quad \alpha > 0, \beta \in \mathbf{R}.$$

Budući da je $\left(\frac{\alpha}{2}\right)^n > 0$, na temelju svojstava eksponencijalne funkcije slijedi da će $\boldsymbol{L}(\alpha,\beta)$ poprimiti najveću vrijednost za ono $\beta \in \mathbf{R}$, za koje izraz $\sum_{i=1}^{n} |x_i - \beta|$ poprimi najmanju vrijednost. No, u II.2. dokazano je da $\sum_{i=1}^{n} |x_i - \beta|$ poprima najmanju vrijednost za $\beta = \hat{m}$, gdje je $\hat{m}$ medijan danog niza podataka $x_1, \ldots, x_n$. *Uzorački medijan* može se, dakako, shvatiti i kao određena s.v., odnosno statistika $\widehat{M}$, i kao takva će biti ML-procjenitelj parametra lokacije β Laplaceove razdiobe.

Deriviranjem jednadžbe (65) po α dobiva se

$$\frac{\partial \boldsymbol{L}(\alpha,\beta)}{\partial \alpha} = \left(\frac{\alpha}{2}\right)^{n-1} \exp\left(-\alpha \sum_{i=1}^{n} |x_i - \beta|\right)\left(\frac{n}{2} - \frac{\alpha}{2}\sum_{i=1}^{n} |x_i - \beta|\right),$$

pa iz $\frac{\partial \boldsymbol{L}(\alpha,\beta)}{\partial \alpha} = 0$ odmah proizlazi

(66) $$\alpha = \frac{n}{\sum_{i=1}^{n} |x_i - \beta|},$$

na temelju čega se zaključuje da je ML-procjenitelj parametra oblika α Laplaceove razdiobe statistika

$$\hat{A} = \frac{n}{\sum_{i=1}^{n} |X_i - \widehat{M}|}.$$

Stavi li se $\alpha_0 = \frac{1}{\alpha}$, onda (66) postaje $\alpha_0 = \frac{1}{n}\sum_{i=1}^{n} |x_i - \beta|$, pa se statistika

(67) $$\hat{A}_0 = \frac{1}{n}\sum_{i=1}^{n} |X_i - \widehat{M}|$$

pojavljuje kao ML-procjenitelj za parametar α_0.

Zanimljivo je uočiti da se formula (67) formalno podudara s formulom (18) iz II.4, tako da se statistika $\hat{A}_0$, definirana u (67) zove *uzoračka apsolutna devijacija oko medijana*.

Usporede li se procjenitelji za parametre α i β Laplaceove razdiobe dobiveni metodom momenata (formule (64)) i ove dobivene metodom najveće vjerojatnosti, vidi se da su to posve različiti procjenitelji. U prvom slučaju procjenitelji se zasnivaju na uzoračkoj aritmetičkoj sredini $\overline{X}$ i uzoračkoj varijanci $\hat{\Sigma}^2$, dok se u drugom slučaju zasnivaju na uzoračkom medijanu $\widehat{M}$ i uzoračkoj apsolutnoj devijaciji oko medijana $\hat{A}_0$. Da bi se utvrdilo koji su procjenitelji bolji, trebalo bi naći odgovarajuće funkcije rizika. Međutim, nalaženje funkcije rizika za procjenitelje $\widehat{M}$ i $\hat{A}_0$ vrlo je složeno, tako da se funkcija rizika ne može izraziti jednostavnom formulom.

Općenito je poznato da su ML-procjenitelji, ako postoje, jednako dobri ili bolji od procjenitelja dobivenih metodom momenata, posebno za velike uzorke ($n \to \infty$).

6. Invarijantnost

Jedno vrlo korisno opće svojstvo ML-procjenitelja, obično se zove *svojstvo invarijantnosti*, sastoji se u sljedećem : Ako je $\hat{T}$ ML-procjenitelj za parametar t, tada je $h(\hat{T})$ ML-procjenitelj za vrijednost $h(t)$, gdje je h određena realna funkcija definirana na skupu Θ ($\Theta \subseteq \mathbf{R}$) dopuštenih vrijednosti parametra t. Odmah se, dakako, mogu postaviti i pitanja o odnosu drugih svojstava procjenitelja $\hat{T}$ i $h(\hat{T})$. Tako se, na primjer, prirodno postavlja pitanje da li iz nepristranosti procjenitelja $\hat{T}$, kao procjenitelja za nepoznati parametar t, proizlazi i nepristranost procjenitelja $h(\hat{T})$, kao procjenitelja za parametar $h(t)$. Također, nužno je odgovoriti na pitanje o odnosu pripadnih funkcija rizika. U tom pogledu ne postoje neki značajniji općeniti rezultati, ali se pitanje odnosa pripadnih funkcija rizika može približno riješiti uz vrlo općenite pretpostavke.

Ako je h derivabilna funkcija i $\hat{t}$ blizu t, onda se na temelju poznatog teorema srednje vrijednosti može pisati

$$h(\hat{t}) - h(t) \approx h'(t)(\hat{t} - t),$$

gdje h' označuje derivaciju funkcije h. Iz toga proizlazi da je

(68) $$\mathrm{E}[(h(\hat{T}) - h(t))^2] \approx (h'(t))^2 \mathrm{E}[(\hat{T} - t)^2] = (h'(t))^2 R(t), t \in \Theta,$$

gdje je $t \mapsto R(t)$ funkcija rizika pri procjeni nepoznatog parametra t procjeniteljem $\hat{T}$. Relacija (68) omogućuje da se dobije uvid u veličinu greške (očekivanu kvadratnu grešku) pri procjeni nepoznate veličine $h(t)$ procjeniteljem $h(\hat{T})$.

Ako se želi dobiti uvid u nepristranost pri procjenjivanju nepoznate veličine $h(t)$ procjeniteljem $h(\hat{T})$, mora se pretpostaviti da je h bar dva puta derivabilna funkcija. Aproksimacijom funkcije h u okolini točke t s prva tri člana pripadnog Taylorova polinoma dobiva se

$$h(x) \approx h(t) + h'(t)(x - t) + \frac{1}{2}h''(t)(x - t)^2.$$

Stavljanjem $x = \widehat{T}$ i uzimanjem očekivanja lijeve i desne strane, dobiva se

$$\mathrm{E}[h(\widehat{T})] \approx h(t) + h'(t)\mathrm{E}[\widehat{T} - t] + \frac{1}{2}h''(t)\mathrm{E}[\widehat{T} - t)^2].$$

Budući da je $\mathrm{E}[\widehat{T} - t] = \mathrm{E}[\widehat{T}] - t = b(t)$ pristranost procjenitelja $\widehat{T}$ pri procjeni parametra t, dok je $\mathrm{E}[(\widehat{T} - t)^2] = R(t)$ odgovarajuća očekivana kvadratna greška, naposljetku se dobiva

$$\text{(69)} \qquad \mathrm{E}[h(\widehat{T})] \approx h(t) + h'(t)b(t) + \frac{1}{2}h''(t)R(t), t \in \Theta.$$

Iz (69) može se zaključiti kada će $h(\widehat{T})$ biti bar približno nepristrani procjenitelj za $h(t)$. Slučaj $h'(t) = 0$ i $R(t) = 0$ očigledno nije zanimljiv. Ako je ispunjen uvjet $b(t) = 0$ i $h''(t) = 0$, onda će također biti $\mathrm{E}[h(\widehat{T})] \approx h(t)$, što znači da će $h(\widehat{T})$ biti približno nepristrani procjenitelj za $h(t)$, ako je $\widehat{T}$ nepristrani procjenitelj za t i graf funkcije $t \longmapsto h(t)$ ima zakrivljenost nula, tj. riječ je o pravcu.

Svojstvo invarijantnosti omogućuje, na primjer, da se ustanovi da je statistika

$$\widehat{\Sigma} = \sqrt{\widehat{\Sigma}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X)^2}$$

ML-procjenitelj za standardnu devijaciju σ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$. Ovdje je, naime, $t = \sigma^2$ i $h(t) = \sqrt{t} = \sigma$, pa kako je $\widehat{\Sigma}^2$ ML-procjenitelj za σ^2, invarijantnost nam jamči da je $\widehat{\Sigma} = h(\widehat{\Sigma}^2) = \sqrt{\widehat{\Sigma}^2}$ ML-procjenitelj za σ. Primijeni li se formula (68) radi približnog dobivanja odgovarajuće funkcije rizika i pozivajući se na formulu (48) dobiva se

$$\mathrm{E}[(\widehat{\Sigma} - \sigma)^2] \approx \frac{2n - 1}{4n^2}\sigma^3, \sigma > 0,$$

što omogućuje određeni uvid u očekivanu kvadratnu grešku pri procjeni standardne devijacije normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$ procjeniteljem $\widehat{\Sigma}$. Primjeni li se, pak, formula (69), dobiva se

$$\mathrm{E}[\widehat{\Sigma}] \approx \sigma - \frac{1}{n}\left(1 - \frac{1}{4n}\right)\sigma,$$

iz čega se razabire da je pristranost procjenitelja $\widehat{\Sigma}$ pri procjeni standardne devijacije σ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$ približno jednaka

$$b(\sigma) \approx -\frac{1}{n}\left(1 - \frac{1}{4n}\right)\sigma, \quad \sigma > 0.$$

U mnogim praktičnim situacijama potrebno je procijeniti nepoznatu vjerojatnost $P_t(I) = \mathrm{P}(a < X \leq b), a < b$, da promatrana slučajna varijabla X poprimi vrijednost iz intervala $I = \langle a, b]$. Očigledno je $P_t(I)$ određena funkcija parametra t, pa se može pisati $P_t(I) = h(t), t \in \Theta$. Ako se raspolaže ML-procjeniteljem $\widehat{T}$ za nepoznati parametar t, onda je $h(\widehat{T})$, na temelju svojstva invarijantnosti, ML-procjenitelj za nepoznatu vjerojatnost $P_t(I)$. Stavi li se $a = -\infty$ i $b = x$, vidi se da je P_t pripadna razdioba vjerojatnosti. Tada je $h(\widehat{T}) = F_{\widehat{T}}(x)$ ML-procjenitelj za vrijednost $F_t(x)$ funkcije razdiobe vjerojatnosti slučajne varijable X u točki $x \in \mathbf{R}$.

Tako, na primjer, ako se želi naći ML-procjenitelj za vjerojatnost da vijek trajanja X nekoga tehničkog uređaja ne bude veći od x ($x > 0$) vremenskih jedinica, uz pretpostavku $X \sim \mathrm{Ex}\left(\frac{1}{\alpha_0}\right)$, tada je riječ o tome da se procjeni

$$h(\alpha_0) = \mathrm{P}(X \leq x) = F_{\alpha_0}(x) = 1 - \exp\left(-\frac{x}{\alpha_0}\right).$$

U 6. primjeru izvedeno je da je statistika $\widehat{A}_0 = \overline{X}$ ML-procjenitelj za parametar α_0, tako da se na temelju svojstva invarijantnosti zaključuje da je statistika

$$h(\widehat{A}_0) = 1 - \exp\left(-\frac{x}{\widehat{A}_0}\right) = 1 - \exp\left(-\frac{x}{\overline{X}}\right)$$

ML-procjenitelj za vrijednost f.r.v. $F_{\alpha_0}(x)$. Budući da je

$$h'(\alpha_0) = -\frac{x}{\alpha_0^2}\exp\left(-\frac{x}{\alpha_0}\right),$$

iz (44) i (68) proizlazi da je

$$\mathrm{E}[(h(\widehat{A}_0) - h(\alpha_0))^2] \approx \frac{x^2}{n\alpha_0^2}\exp\left(-\frac{2x}{\alpha_0}\right)$$

očekivana kvadratna greška pri procjeni vjerojatnosti $\mathrm{P}(X \leq x)$ procjeniteljem $1 - \exp\left(-\frac{x}{\overline{X}}\right)$.

U 6. primjeru također je pokazano da je $\overline{X}$ nepristrani procjenitelj za parametar α_0, a primjenom formule (69) dobiva se

$$\mathrm{E}\left[1 - \exp\left(-\frac{x}{\overline{X}}\right)\right] \approx 1 - \exp\left(-\frac{x}{\alpha_0}\right) - \frac{x}{2n\alpha_0}\left(\frac{x}{\alpha_0} - 2\right)\exp\left(-\frac{x}{\alpha_0}\right),$$

što pokazuje da statistika $1 - \exp\left(-\frac{x}{\overline{X}}\right)$, kao procjenitelj za nepoznatu vjerojatnost $\mathrm{P}(X \leq x)$, približno ima pristranost

$$b \approx -\frac{x}{2n\alpha_0}\left(\frac{x}{\alpha_0} - 2\right)\exp\left(-\frac{x}{\alpha_0}\right),$$

iz čega se vidi da za $x = 2\alpha_0$ imamo približno nepristrani procjenitelj.

Korisno je uočiti da u izvedenim formulama za pristranost imamo faktor $\frac{1}{n}$, što znači da za velike n ($n \to \infty$) praktički imamo nepristrane procjenitelje.

Svojstvo invarijantnosti ML-procjenitelja vrijedi, dakako, i u višeparametarskom modelu, tj. kada je $\boldsymbol{t} = (t_1, \dots, t_k)$. Ako su, dakle, $\widehat{T}_1, \dots, \widehat{T}_k$ ML-procjenitelji za nepoznate parametre $t_1, \dots, t_k$, tada je $h(\widehat{T}_1, \dots, \widehat{T}_k)$, gdje je h određena realna funkcija k realnih varijabli, ML-procjenitelj za nepoznatu vrijednost $h(t_1, \dots, t_k)$.

Tako, na primjer, ako je riječ o uniformnoj razdiobi $U(a,b)$ na intervalu $\langle a, b\rangle$, ML-procjenitelji za parametre a i b su $\widehat{A} = \min(X_1, \dots, X_n)$ i $\widehat{B} = \max(X_1, \dots, X_n)$ (v. zad. 11). Ako nas, pak, zanima procjena za parametar $m = \frac{a+b}{2}$ (sredina intervala), ili recimo za parametar $l = b - a$ (širina intervala), onda nam svojstvo invarijantnosti jamči da su $\widehat{M} = \frac{\widehat{A}+\widehat{B}}{2}$ i $\widehat{L} = \widehat{B} - \widehat{A}$ odgovarajući ML-procjenitelji za parametre m i l.

7. Efikasnost

Opći kriterij za vrednovanje različitih procjenitelja istoga nepoznatog parametra sadržan je u minimaks-principu (v. VI.1), koji se za nepristrane procjenitelje svodi na konstataciju da je najbolji onaj procjenitelj koji ima najmanju varijancu. Stoga je logično da se kaže da je nepristrani procjenitelj $\widehat{T}_1$ ***efikasniji*** od nepristranog procjenitelja $\widehat{T}_2$ za nepoznati parametar t, ako vrijedi

$$\text{(70)} \qquad \mathrm{V}[\widehat{T}_1] < \mathrm{V}[\widehat{T}_2], \quad \forall t \in \Theta.$$

Drugim riječima, pri procjeni nepoznatog parametra t vrijednošću $\hat{t}_1$ efikasnijeg procjenitelja $\widehat{T}_1$ očekuje se manja kvadratna greška nego pri procjeni vrijednošću $\hat{t}_2$ manje efikasnog procjenitelja $\widehat{T}_2$. Stoga se prirodno nameće zadatak da se pronađe (ako postoji) *najefikasniji procjenitelj*, tj. onaj kojemu pripada najmanja varijanca, ili bar da se odredi donja granica za vrijednosti varijanci svih mogućih nepristranih procjenitelja za parametar t.

Da bi se riješio taj zadatak treba uvesti još neke dodatne pretpostavke u već opisani teorijski model. Vjerojatnosna razdioba P_t u klasi $\mathcal{P} = \{P_t : t \in \Theta\}$ dopuštenih razdioba obično je diskretna ili kontinuirana razdioba vjerojatnosti.

Kontinuirana razdioba vjerojatnosti zadana je svojom f.g.v. $x \mapsto f_t(x), x \in \mathbf{R}$. No, za fiksirano $x \in \mathbf{R}$ može se promatrati i funkcija $t \mapsto f_t(x), t \in \Theta$, pa se zahtijeva da ta funkcija bude "dovoljno glatka", tj. da postoji neprekidna druga derivacija $\frac{\partial^2 f_t(x)}{\partial t^2}$, te da je dopušteno derivirati pod znakom integrala. Tada se, naime, može definirati izraz

$$\text{(71)} \qquad I(t) = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial t} \ln f_t(x)\right]^2 f_t(x)\,\mathrm{d}\,x = \mathrm{E}\left[\left(\frac{\partial}{\partial t} \ln f_t(X)\right)^2\right],$$

pri čemu se $I(t)$ zove ***Fisherova informacija*** vjerojatnosne razdiobe P_t.

Ako je riječ o diskretnoj vjerojatnosnoj razdiobi P_t, onda se Fisherova informacija definira formulom

$$\text{(72)} \qquad I(t) = \sum_j \left[\frac{\partial}{\partial t} \ln P_t(a_j)\right]^2 P_t(a_j) = \mathrm{E}\left[\left(\frac{\partial}{\partial t} \ln P_t(X)\right)^2\right],$$

gdje je $A = \{a_j \in \mathbf{R} : j = 1, 2, \dots\}$ skup vrijednosti za klasu diskretnih razdioba P_t, odnosno skup vrijednosti diskretne s.v. X (v. IV.1). Pritom je $P_t(a_j) = \mathrm{P}(X = a_j)$ i također se pretpostavljaju odgovarajući uvjeti za funkciju $t \mapsto P_t(a_j), t \in \Theta$.

Sada se može izreći i čuvena ***Rao-Cramerova nejednakost***, koja kaže da za svaki nepristrani procjenitelj $\widehat{T}$ parametra t, u opisanom teorijskom modelu, vrijedi

$$\text{(73)} \qquad V[\widehat{T}] \geq \frac{1}{nI(t)}, \quad t \in \Theta,$$

gdje je n veličina uzorka.

Veličina $\frac{1}{nI(t)}$ zove se *Rao-Cramerova donja granica* za varijance nepristranih procjenitelja.

Dokažimo Rao-Cramerovu nejednakost kod kontinuiranih razdioba. Neka je $\widehat{T} = h(X_1, \dots, X_n)$ nepristrani procjenitelj za parametar t. Na temelju definicije matematičkog očekivanja i pojma nepristranosti može se pisati

$$\mathrm{E}[\widehat{T}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) f_t(x_1) \cdot \ldots \cdot f_t(x_n)\,\mathrm{d}\,x_1 \cdot \ldots \cdot \mathrm{d}\,x_n = t,$$

pri čemu se, dakako, imalo na umu da su $X_1, \dots, X_n$ nezavisne slučajne varijable. Deriviranjem te jednadžbe po t dobiva se

$$\text{(74)} \quad \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \dots, x_n) \left[\sum_{i=1}^{n} \frac{\partial}{\partial t} \ln f_t(x_i)\right] f_t(x_1) \cdot \ldots \cdot f_t(x_n)\,\mathrm{d}\,x_1 \cdot \ldots \cdot \mathrm{d}\,x_n = 1,$$

pri čemu se imalo na umu da se derivacija produkta može izraziti formulom

$$\text{(75)} \qquad \frac{\partial}{\partial t}[f_t(x_1) \cdot \ldots \cdot f_t(x_n)] = \left[\sum_{i=1}^{n} \frac{1}{f_t(x_i)} \frac{\partial}{\partial t} f_t(x_i)\right] f_t(x_1) \cdot \ldots \cdot f_t(x_n),$$

te da se derivacija logaritma može zapisati

$$\text{(76)} \qquad \frac{\partial}{\partial t}[\ln f_t(x_i)] = \frac{1}{f_t(x_i)} \frac{\partial}{\partial t} f_t(x_i), \quad i = 1, \dots, n.$$

Stavimo $Z_i = \frac{\partial}{\partial t}[\ln f_t(X_i)]$, $i = 1, \dots, n$, pa se vidi da su $Z_1, \dots, Z_n$ nezavisne s

Ako se još stavi $Z = \sum_{i=1}^{n} Z_i = g(X_1, \ldots, X_n)$, jednadžba (74) može se zapisati u obliku

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) g(x_1, \ldots, x_n) f_t(x_1) \cdot \ldots \cdot f_t(x_n) \, \mathrm{d} x_1 \cdot \ldots \cdot \mathrm{d} x_n = 1,$$

odnosno, primjenom operatora očekivanja E može se zapisati u obliku

(77) $$\mathrm{E}[h(X_1, \ldots, X_n) g(X_1, \ldots, X_n)] = \mathrm{E}[\widehat{T} Z] = 1.$$

Polazeći, pak, od očigledne jednakosti $\int_{-\infty}^{\infty} f_t(x_i) \, \mathrm{d} x_i = 1, i = 1, \ldots, n$, deriviranjem po t i uzimanjem u obzir (76), dobiva se

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial t} [\ln f_t(x_i)] \, f_t(x_i) \, \mathrm{d} x_i = 0, \quad i = 1, \ldots, n,$$

što se može zapisati i kao $\mathrm{E}[Z_i] = 0$. Iz toga slijedi da je $\mathrm{E}[Z] = 0$ i da je

(78) $$\mathrm{V}[Z] = \mathrm{V}\left[\sum_{i=1}^{n} Z_i\right] = \sum_{i=1}^{n} \mathrm{E}\left[Z_i^2\right] = n \mathrm{E}\left[\left(\frac{\partial}{\partial t} \ln f_t(X)\right)^2\right].$$

Iz definicije kovarijance (v. (30) u V.4) proizlazi da se može pisati

(79) $$\mathrm{Cov}(\widehat{T}, Z) = \mathrm{E}[\widehat{T} Z] - \mathrm{E}[\widehat{T}] \, \mathrm{E}[Z].$$

Budući da je $\mathrm{E}[Z] = 0$, iz (77) i (79) slijedi

(80) $$\mathrm{Cov}(\widehat{T}, Z) = 1.$$

Primjenjujući na slučajne varijable $\widehat{T}$ i $\widehat{Z}$ poznati rezultat da kvadrat koeficijenta korelacije ne premašuje jedan (v. V.4), može se pisati

$$\frac{(\mathrm{Cov}(\widehat{T}, Z))^2}{\mathrm{V}[\widehat{T}] \mathrm{V}[Z]} \leq 1,$$

što s obzirom na (71), (78) i (80) postaje (73), tj. Rao-Cramerova nejednakost.

Slično se može izvesti Rao-Cramerova nejednakost i za diskretne vjerojatnosne razdiobe P_t $(t \in \Theta)$ (v. zad. 23).

Sada je jasno da će nepristrani procjenitelj $\widehat{T}_0$, za koji vrijedi

(81) $$\mathrm{V}[\widehat{T}_0] = \frac{1}{n I(t)}, \quad t \in \Theta,$$

biti najefikasniji procjenitelj za parametar t u klasi svih nepristranih procjenitelja.

8. primjer

Neka je, kao i u 6. primjeru riječ o procjeni parametra α eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$. Vidjelo se, zapravo, da je prikladnije promatrati parametar $\alpha_0 = \frac{1}{\alpha}$, i tada se pokazalo da je $\widehat{A}_0 = \overline{X}$ nepristrani ML-procjenitelj za parametar α_0, uz funkciju rizika $\alpha_0 \mapsto R(\alpha_0) = \mathrm{V}[A_0] = \frac{1}{n} \alpha_0^2$ $(\alpha_0 > 0)$. Ako se, pak, izračuna, prema (71), odgovarajuća Fisherova informacija, dobiva se

$$I(\alpha_0) = \mathrm{E}\left[\left(\frac{\partial}{\partial \alpha_0} \ln \left(\frac{1}{\alpha_0} \exp\left(-\frac{x}{\alpha_0}\right)\right)\right)^2\right] = \frac{1}{\alpha_0^2},$$

iz čega se razabire da je Rao-Cramerova donja granica

$$\frac{1}{n I(\alpha_0)} = \frac{\alpha_0^2}{n} = \mathrm{V}[\widehat{A}_0],$$

što znači da je $\widehat{A}_0 = \overline{X}$ najefikasniji nepristrani procjenitelj za nepoznati parametar $\alpha_0 = \frac{1}{\alpha}$ eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$.

9. primjer

U 5. primjeru nađeno je da je $\widehat{\Lambda} = \overline{X}$ nepristrani ML-procjenitelj za parametar λ $(\lambda > 0)$ Poissonove razdiobe $\mathrm{Po}(\lambda)$. Također je određena i odgovarajuća funkcija rizika $\lambda \mapsto R(\lambda) = \frac{1}{n} \lambda, \lambda > 0$. Da bi se pokazalo da je $\widehat{\Lambda}$ i najefikasniji nepristrani procjenitelj za parametar λ, treba izračunati odgovarajuću Fisherovu informaciju. Prema (72) dobiva se

$$I(\lambda) = \mathrm{E}\left[\left(\frac{\partial}{\partial \lambda} \ln \left(\frac{\lambda^X}{X!} \exp(-\lambda)\right)\right)^2\right] = \mathrm{E}\left[\left(\frac{X - \lambda}{\lambda}\right)^2\right] = \frac{1}{\lambda^2} \mathrm{V}[X].$$

Budući da je za Poissonovu razdiobu $\mathrm{V}[X] = \lambda$ (v. IV.3), dobiva se

$$I(\lambda) = \frac{1}{\lambda},$$

iz čega odmah slijedi da je Rao-Cramerova donja granica

$$\frac{1}{n I(\lambda)} = \frac{1}{n} \lambda = \mathrm{V}[\widehat{\Lambda}],$$

što pokazuje da je $\widehat{\Lambda} = \overline{X}$ najefikasniji nepristrani procjenitelj za parametar λ Poissonove razdiobe $\mathrm{Po}(\lambda)$.

10. primjer

U VI.3. izvedeno je da je statistika $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ nepristrani procjenitelj za parametar σ^2 normalne razdiobe $N(\mu, \sigma^2)$. Stavljajući $\sigma^2 = t$ i polazeći od f.g.v. $f_t(x) = \frac{1}{\sqrt{2\pi t}} \exp\left[-\frac{(x-\mu)^2}{2t}\right]$ normalne razdiobe $N(\mu, t)$, primjenom formule (71), dobiva se pripadna Fisherova informacija

$$I(t) = \mathrm{E}\left[\left(\frac{\partial}{\partial t}\left(-\frac{1}{2}\ln\sqrt{2\pi} - \frac{1}{2}\ln t - \frac{(X-\mu)^2}{2t}\right)\right)^2\right] =$$

$$= \mathrm{E}\left[\left(-\frac{1}{2t} + \frac{(X-\mu)^2}{2t^2}\right)^2\right] =$$

$$= \frac{1}{4t^2}\mathrm{E}[(X-\mu)^4] - \frac{1}{2t^3}\mathrm{E}[(X-\mu)^2] + \frac{1}{4t^2}.$$

Budući da je u normalnoj razdiobi $N(\mu, \sigma^2)$ četvrti centralni moment $\mu_4 = \mathrm{E}[(X-\mu)^4] = 3\sigma^4 = 3t^2$, dok je $\mathrm{E}[(X-\mu)^2] = \mathrm{V}[X] = \sigma^2 = t$, naposljetku se dobiva

$$I(t) = \frac{1}{2t^2}, t > 0,$$

tako da Rao-Cramerova donja granica glasi

$$\frac{1}{nI(t)} = \frac{2t^2}{n} = \frac{2}{n}\sigma^4.$$

Uspoređujući to s varijancom $\mathrm{V}[S^2] = \frac{2}{n-1}\sigma^4$, odmah se vidi da za $n > 1$ nepristrani procjenitelj S^2 parametra σ^2 ne postiže Rao-Cramerovu donju granicu, što upućuje na zaključak da S^2 nije najefikasniji procjenitelj za nepoznatu varijancu normalne razdiobe $N(\mu, \sigma^2)$.

Općenito se funkcija

(82) $$e(t) = \frac{1}{n\mathrm{V}[\widehat{T}]I(t)}, t \in \Theta,$$

definirana kao omjer između Rao-Cramerove donje granice i varijance procjenitelja $\widehat{T}$ zove ***efikasnost*** procjenitelja $\widehat{T}$.

Za najefikasnije procjenitelje očigledno je efikasnost jednaka jedinici, dok općenito vrijedi

$$0 \leq e(t) \leq 1.$$

Efikasnost procjenitelja S^2 pri procjeni nepoznate varijance σ^2 normalne razdiobe $N(\mu, \sigma^2)$ iznosi $e(t) = e(\sigma^2) = \frac{n-1}{n}$, iz čega se vidi da efikasnost ne ovisi o parametru σ^2 i da je za velike uzorke ($n \to \infty$) ta efikasnost vrlo bliska jedinici.

8. Asimptotska svojstva procjenitelja

U dosadašnjim razmatranjima procjenitelj nepoznatog parametra razmatran je kao određena statistika, tj. funkcija fiksiranog broja n ($n \in \mathbf{N}$) nezavisnih slučajnih varijabli $X_1, \ldots, X_n$ sa zajedničkom razdiobom vjerojatnosti. Pretpostavljalo se, dakle, da je veličina n slučajnog uzorka $(X_1, \ldots, X_n)$ konstantna veličina. Očito je, međutim, da u primjenama teorije statističkog zaključivanja veličina uzorka može varirati. U nekim situacijama procjena nepoznatog parametra donosit će se na temelju malog broja mjerenja, odnosno opažanja, dok će u drugim okolnostima istraživač raspolagati velikim brojem mjerenja. S druge, pak, strane, pokazat će se da mnoga svojstva procjenitelja, na kojima se temelji praktična primjena teorije statističkog zaključivanja, bitno ovise o veličini uzorka. Postupak mjerenja, odnosno opažanja, kojim se dolazi do statističkih podataka $x_1, \ldots, x_n$, redovito je povezan s odgovarajućim troškovima ovisnima o broju mjerenja n, tako da sve to upućuje na potrebu da se istraže svojstva procjenitelja s obzirom na variranje veličine slučajnog uzorka n.

U tu svrhu uvodi se, za procjenitelja nepoznatog parametra t, oznaka

(83) $$\widehat{T}_n = h(X_1, \ldots, X_n), n \in \mathbf{N},$$

da bi se istakla i ovisnost o n. Relacijom (83) definiran je, zapravo, beskonačni niz slučajnih varijabli $\widehat{T}_1, \widehat{T}_2, \ldots, \widehat{T}_n, \ldots$, pa se prirodno nameće ideja da se razmotre pitanja u vezi s konvergencijom toga niza. Rezultati koji se tako dobiju obično se zovu *asimptotskim svojstvima procjenitelja*.

Primijetimo najprije da gotovo sve izvedene formule za fukciju rizika sadrže n (v. npr. (5), (9), (17), (22), (23), itd.), pa se i za funkciju rizika uvodi oznaka

(84) $$R_n(t) = \mathrm{E}[(\widehat{T}_n - t)^2], t \in \Theta, n \in \mathbf{N}.$$

Za svaki $t \in \Theta$ formulom (84) definiran je beskonačni niz realnih brojeva, pa ako taj niz konvergira k nuli za svaki $t \in \Theta$, tj. vrijedi

(85) $$\lim_{n\to\infty} \mathrm{E}[(\widehat{T}_n - t)^2] = 0, t \in \Theta,$$

kaže se da je $\widehat{T}_n$ ***konzistentan procjenitelj*** za parametar t.

Svojstvo konzistentnosti jamči, zapravo, da se srednja kvadratna greška pri aproksimaciji nepoznatog parametra t vrijednošću $\hat{t}_n$ konzistentnog procjenitelja $\widehat{T}_n$ može učiniti po volji malenom ako se se samo uzme dovoljno velik broj n mjerenja, odnosno opažanja promatrane slučajne varijable X. Ono također pokazuje da bi se s beskonačno velikim uzorkom dobila točna procjena. Naravno da je u praksi nemoguće postići beskonačno velik uzorak, ali je također jasno da je konzistentnost vrlo poželjno svojstvo.

Gotovo svi razmotreni procjenitelji imaju svojstvo konzistentnosti. Da postoje i nekonzistentni procjenitelji pokazuje relacija (9), iz koje se vidi da je za $a \neq$ $\neq 1$, $\lim_{n\to\infty} R_1(p) = \infty$, za svaki $p \neq 0$, a to znači da procjenitelj $a\overline{X}$ ($a \neq 1$) nije

konzistentan procjenitelj za parametar p, binomne razdiobe $B(1,p)$. Za $a = 1$ riječ je o procjenitelju $\overline{X}$ i on je, dakako, konzistentan procjenitelj za p, što se vidi i iz relacije (5), iz koje proizlazi da je

$$\lim_{n\to\infty} R(p) = \lim_{n\to\infty} \frac{1}{n} p(1-p) = 0, \quad p \in [0,1].$$

Pojam konzistentnosti definira se ponegdje i drukčije. Kaže se, naime, da je $\widehat{T}_n$ ($n \in \mathbf{N}$) konzistentan procjenitelj za parametar t, ako za svaki $\delta > 0$ vrijedi

(86) $$\lim_{n\to\infty} \mathrm{P}(|\widehat{T}_n - t| \geq \delta) = 0, \quad t \in \Theta.$$

Tom definicijom konzistentnost se opisuje pomoću vjerojatnosti da apsolutna razlika između procjenitelja i nepoznatog parametra premaši proizvoljno maleni pozitivni broj δ, za koju (vjerojatnost) se zahtijeva da teži nuli kada veličina uzorka n teži u beskonačnost.

Formulom (86) je, inače, definiran pojam ***stohastičke konvergencije***, ili ***konvergencije po vjerojatnosti***, niza slučajnih varijabli $\widehat{T}_1, \widehat{T}_2, \ldots, \widehat{T}_n, \ldots$ neslučajnoj veličini t. To se zapisuje

(87) $$\lim_{n\to\infty} \mathrm{st}\, \widehat{T}_n = t,$$

pa se može reći da konzistentan procjenitelj stohastički konvergira parametru koji procjenjuje. Na temelju Čebiševljeve nejednakosti (v. (35) u IV.4) proizlazi da je

$$\mathrm{E}[(\widehat{T}_n - t)^2] \geq \delta^2 \mathrm{P}(|\widehat{T}_n - t| \geq \delta),$$

što pokazuje da iz (85) slijedi (86), a to znači da je svaki konzistentni procjenitelj po prvoj definiciji, konzistentan i po drugoj. Obratno općenito ne vrijedi i stoga ćemo se redovito oslanjati na prvu definiciju, izraženu formulom (85).

U VI.7. uveden je pojam efikasnosti procjenitelja (v. (82)), a vidjelo se i da neki vrlo značajni procjenitelji, kao na primjer S^2 u 10. primjeru, nisu najefikasniji u smislu da im efikasnost iznosi jedan. Izvedeno je, naime, da je $e(\sigma^2) = \dfrac{n-1}{n}$, iz čega se razabire da efikasnost ovisi o veličini uzorka n, pa je prirodno da se, umjesto $e(\sigma^2)$, piše $e_n(\sigma^2)$ i promotri

$$\lim_{n\to\infty} e_n(\sigma^2) = \lim_{n\to\infty} \frac{n-1}{n} = 1,$$

što pokazuje da je, u asimptotskom smislu ($n \to \infty$), efikasnost procjenitelja S^2 jednaka jedan.

Zato se općenito, u vezi s nizom procjenitelja $\widehat{T}_n$ ($n \in \mathbf{N}$), promatra niz pripadnih efikasnosti $e_n(t)$ ($n \in \mathbf{N}$), i ako postoji

(88) $$\lim_{n\to\infty} e_n(t) = e_0(t), t \in \Theta,$$

onda se funkcija $t \mapsto e_0(t)$ zove ***asimptotska efikasnost*** procjenitelja $\widehat{T}_n$.

Očigledno je $0 \leq e_0(t) \leq 1$ ($t \in \Theta$) i ako je $e_0(t) = 1$ kaže se da je procjenitelj $\widehat{T}_n$ ***asimptotski efikasan***.

Tako se, na primjer, za korigiranu uzoračku varijancu S^2 može reći da je, kao procjenitelj za nepoznatu varijancu σ^2 normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$, asimptotski efikasan procjenitelj, iako je za konačne $n \in \mathbf{N}$ njegova efikasnost manja od jedinice.

Za konzistentne ML-procjenitelje općenito vrijedi da su to asimptotski efikasni procjenitelji, ako već nisu najefikasniji.

Na kraju ćemo razmotriti još jedno asimptotsko svojstvo procjenitelja koje će imati veliku važnost i primjenu u različitim problemima teorije statističkog zaključivanja. Primijetimo najprije da je procjenitelj $\widehat{T}_n$, za svaki $n \in \mathbf{N}$, određena slučajna varijabla kojoj pripada odgovarajuća vjerojatnosna razdioba, općenito ovisna o nepoznatom parametru t, tj. o vjerojatnosnoj razdiobi P_t pretpostavljenoga teorijskog modela.

Tako smo u 1. primjeru imali procjenitelj $\overline{X} = \overline{X}_n$ za nepoznati parametar p binomne razdiobe $B(1,p)$, kojemu je pripadala razdioba vjerojatnosti opisana formulom (4).

Također smo u VI.4. pokazali da uzoračkoj aritmetičkoj sredini $\overline{X}$, kao procjenitelju za nepoznati parametar μ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$, pripada normalna razdioba $\mathrm{N}\left(\mu, \dfrac{1}{n}\sigma^2\right)$.

Općenito je problem pronalaženja vjerojatnosne razdiobe procjenitelja $\widehat{T}_n$, u teorijskom modelu s klasom dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{P_t : t \in \Theta\}$, vrlo složen i samo u nekim posebnim situacijama može se dobiti jednostavno rješenje uz konačni $n \in \mathbf{N}$. Međutim, uz vrlo općenite pretpostavke može se dobiti asimptotsko ($n \to \infty$) rješenje problema. To rješenje temelji se na čuvenom rezultatu iz teorije vjerojatnosti (v. [38]), poznatom pod nazivom ***centralni granični teorem*** (CGT). Jedna od verzija CGT-a izriče slijedeće:

Ako je X_i ($i = 1, 2, \ldots$) niz nezavisnih slučajnih varijabli kojima pripada ista vjerojatnosna razdioba s očekivanjem μ i varijancom σ^2 ($0 < \sigma^2 < \infty$) i $Y_n = \sum_{i=1}^{n} X_i$, tada slučajne varijable

$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}, n \in \mathbf{N},$$

imaju svojstvo da pripadajući niz funkcija razdiobe vjerojatnosti

$$G_n(z) = \mathrm{P}(Z_n \leq z), n \in \mathbf{N}$$

konvergira funkciji $\Phi(z) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\dfrac{1}{2}x^2\right) \mathrm{d}\,x$, tj. funkciji razdiobe vjerojatnosti standardne normalne razdiobe $\mathrm{N}(0,1)$.

Kaže se još da niz slučajnih varijabli Z_n $(n \in \mathbf{N})$ ***konvergira po razdiobi*** slučajnoj varijabli $Z \sim \mathrm{N}(0,1)$ i piše se

$$Z_n \xrightarrow{\mathcal{D}} Z.$$

Za slučajnu varijablu Y_n kaže se, pak, da je ***asimptotski normalna***.

Budući da je $\mathrm{E}[Y_n] = n\mu$ i $\mathrm{V}[Y_n] = n\sigma^2$, odmah se vidi da je $Z_n = \dfrac{Y_n - \mathrm{E}[Y_n]}{\sqrt{\mathrm{V}[Y_n]}}$, pa asimptotska normalnost, zapravo, znači da se za velike n može uzeti da Y_n približno ima normalnu razdiobu $\mathrm{N}(n\mu, n\sigma^2)$.

Sada je jasno kako će se definirati pojam asimptotske normalnosti procjenitelja $\widehat{T}_n$.

> Reći će se da je $\widehat{T}_n$ ***asimptotski normalan procjenitelj*** ako niz slučajnih varijabli
>
> $$\frac{\widehat{T}_n - \mathrm{E}[\widehat{T}_n]}{\sqrt{\mathrm{V}[\widehat{T}_n]}}, n = 1, 2, \dots$$
>
> konvergira po razdiobi standardnoj normalnoj slučajnoj varijabli.

Ako je riječ o nepristranom procjenitelju $\widehat{T}_n$ parametra t, onda je $\mathrm{E}[\widehat{T}_n] = t$ i $\mathrm{V}[\widehat{T}_n] = R_n(t)$, gdje je $R_n(t)$ vrijednost funkcije rizika, pa se može reći da za nepristrani, asimptotski normalan procjenitelj $\widehat{T}_n$ nepoznatog parametra t za velike n približno vrijedi

(89) $$\widehat{T}_n \sim \mathrm{N}(t, R_n(t)).$$

11. primjer

Vidjeli smo da se uzoračka aritmetička sredina $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ vrlo često pojavljuje kao procjenitelj za neki parametar, najčešće za nepoznato očekivanje $\mu = \mathrm{E}[X]$ promatrane slučajne varijable X. Očigledno se može pisati $\overline{X}_n = \dfrac{1}{n} Y_n$, gdje je $Y_n = \sum_{i=1}^{n} X_i$ a $X_1, \dots, X_n$ su nezavisne slučajne varijable kojima pripada ista razdioba vjerojatnosti s očekivanjem μ i varijancom $\mathrm{V}[X] = \sigma^2$. Budući da je $\mathrm{E}[\overline{X}] = \mu$ i $\mathrm{V}[\overline{X}] = \dfrac{1}{n}\sigma^2$, može se pisati

$$\frac{\overline{X}_n - \mathrm{E}[\overline{X}_n]}{\sqrt{\mathrm{V}[\overline{X}_n]}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} = \frac{n\overline{X}_n - n\mu}{\sigma\sqrt{n}} = \frac{Y_n - n\mu}{\sigma\sqrt{n}} = Z_n,$$

iz čega se vidi da niz slučajnih varijabli Z_n $(n \in \mathbf{N})$ zadovoljava uvjete CGT-a, a to upravo znači da je $\overline{X}_n$ asimptotski normalan procjenitelj za nepoznato očekivanje

μ. Može se, prema tome, reći da za velike n približno vrijedi

(90) $$\overline{X}_n \sim N\left(\mu, \frac{1}{n}\sigma^2\right).$$

Koliko je bitna pretpostavka u CGT-u da je Y_n zbroj, a ne neka druga funkcija nezavisnih slučajnih varijabli $X_1, \dots, X_n$, pokazuje nam primjer ML-procjenitelja $\widehat{T}_2 = \max(X_1, \dots, X_n)$ za parametar t uniformne razdiobe $\mathrm{U}(0,t)$, koji je razmatran u 2. i 4. primjeru. Uzme li se, naime, u obzir (21) i zatim promotri niz slučajnih varijabli

$$\frac{\widehat{T}_2 - \mathrm{E}[\widehat{T}_2]}{\sqrt{\mathrm{V}[\widehat{T}_2]}} = \frac{(n+1)\widehat{T}_2 - nt}{t}\sqrt{\frac{n+2}{n}}, \quad n \in \mathbf{N},$$

pokazuje se (v. [16]) da taj niz ne konvergira po razdiobi standardnoj normalnoj slučajnoj varijabli, nego slučajnoj varijabli kojoj pripada tzv. *Weibullova razdioba*. To znači da $\widehat{T}_2$ nije asimptotski normalan procjenitelj.

9. Bayesova metoda

Određivanje procjenitelja za nepoznati parametar na temelju načela najveće vjerojatnosti i načela jednakosti momenata zasniva se samo na informaciji koju o nepoznatom parametru daje niz empirijskih vrijednosti $x_1, \dots, x_n$. Postoje, međutim, određene situacije u praksi kada istraživač raspolaže i određenom apriornom informacijom o nepoznatom parametru, koja se obično temelji na subjektivnim spoznajama istraživača, pa je riječ o tome da se izgradi teorijski model koji će odgovorati takvoj situaciji. Jedan takav model poznat je kao *Bayesova metoda procjene parametara*.

> Osim uobičajene pretpostavke da je riječ o klasi $\mathcal{P} = \{P_t : t \in \Theta\}$ dopuštenih vjerojatnosnih razdioba sa skupom Θ dopuštenih vrijednosti parametra, u Bayesovoj metodi se još pretpostavlja da je poznata tzv. *apriorna razdioba vjerojatnosti* parametra t.

Da bi se lakše shvatilo načelo funkcioniranja Bayesove metode, najprije će se razmotriti jedan primjer.

12. primjer

Neka je, kao i u 1. primjeru riječ o procjeni parametra $t = p$ (proporcija neispravnih proizvoda) na temelju $n = 5$ mjerenja x_1, x_2, x_3, x_4 i x_5, pri čemu se još raspolaže i informacijom da je apriorna vjerojatnosna razdioba parametra t diskretnog tipa i da je zadana tablicom 2.

Nepoznati parametar t razmatra se, zapravo, kao diskretna s.v. T sa skupom vrijednosti $\{t_1 = 0{,}4, t_2 = 0{,}5\}$ i pripadnim vjerojatnostima $P(T = t_1) = q_1 = 0{,}8$ i $P(T = t_2) = q_2 = 0{,}2$.

Tablica 2.

i	1	2
t_i	0,4	0,5
q_i	0,8	0,2

Kako je već rečeno u 1. primjeru, x_i $(i = 1, \ldots, 5)$ može poprimiti vrijednost 0 (proizvod nije neispravan) i vrijednost 1 (proizvod je neispravan), pa stoga statistika $\overline{X} = \frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5)$ označuje relativnu frekvenciju neispravnih proizvoda u slučajnom uzorku veličine $n = 5$.

Uz pretpostavku da je stvarna vrijednost nepoznatog parametra baš t $(0 \leq t \leq \leq 1)$, s.v. $Y = 5\overline{X} = X_1 + X_2 + X_3 + X_4 + X_5$ ima binomnu razdiobu $B(5, t)$. To omogućuje da se izračunaju vrijednosti

$$p_{k/i} = \binom{5}{k} t_i^k (1 - t_i)^{5-k}, \quad i = 1, 2, \ k = 0, 1, 2, 3, 4, 5,$$

pri čemu $p_{k/i}$ označuje uvjetnu vjerojatnost da statistika $\overline{X}$ poprimi vrijednost $\frac{k}{5}$, uz uvjet da nepoznati parametar ima vrijednost t_i. Rezultati proračuna prikazani su u tabl. 3.

Tablica 3.

t_i \ k	0	1	2	3	4	5
0,4	0,078	0,259	0,345	0,231	0,077	0,010
0,5	0,031	0,156	0,313	0,313	0,156	0,031

Budući da je simultana vjerojatnost da s.v. $\overline{X}$ poprimi vrijednost $\frac{k}{5}$ i s.v. T vrijednost t_i izražena formulom

$$(91) \quad P(\overline{X} = \frac{k}{5}, T = t_i) = P(T = t_i)P(\overline{X} = \frac{k}{5}/T = t_i) = q_i p_{k/i}, \qquad i = 1, 2, \ k = 0, 1, 2, 3, 4, 5,$$

bit će

$$(92) \quad P(\overline{X} = \frac{k}{5}) = q_1 p_{k/1} + q_2 p_{k/2}, \quad k = 0, 1, 2, 3, 4, 5.$$

Sada se može izračunati i uvjetna vjerojatnost da nepoznati parametar poprimi vrijednost t_i uz uvjet da je statistika $\overline{X}$ poprimila vrijednost $\frac{k}{5}$, tj. može se pisati

$$(93) \quad P(T = t_i/\overline{X} = \frac{k}{5}) = q_{i/k} = \frac{P(\overline{X} = \frac{k}{5}, T = t_i)}{P(\overline{X} = \frac{k}{5})}, \quad i = 1, 2, \ k = 0, 1, 2, 3, 4, 5.$$

Iz tabl. 2. i 3. primjenom formula (91), (92) i (93), dobiva se tabl. 4, koja sadrži vrijednosti $q_{i/k}$.

Tablica 4.

t_i \ k	0	1	2	3	4	5
0,4	0,912	0,870	0,816	0,748	0,665	0,571
0,5	0,088	0,130	0,184	0,252	0,335	0,429

U tablici 4. prikazana je, zapravo, informacija o nepoznatom parametru sadržana u apriornoj vjerojatnosnoj razdiobi i u statistici $\overline{X}$, za koju smo već ranije utvrdili da je dobar procjenitelj za nepoznati parametar t. Ako se, na primjer, za statistiku $\overline{X}$ dobije vrijednost $\overline{x} = \frac{1}{5}$ $(k = 1)$, onda tabl. 4. sugerira da se, umjesto apriorne razdiobe, za nepoznati parametar usvoji *aposteriorna razdioba*

$$q_{1/1} = P\left(T = 0{,}4/\overline{X} = \frac{1}{5}\right) = 0{,}870, \qquad q_{2/1} = P\left(T = 0{,}5/\overline{X} = \frac{1}{5}\right) = 0{,}130.$$

Matematičko očekivanje te razdiobe, tj. uvjetno očekivanje s.v. T uz uvjet da je $\overline{X} = \frac{1}{5}$, iznosi

$$E\left[T/\overline{X} = \frac{1}{5}\right] = t_1 q_{1/1} + t_2 q_{2/1} = 0{,}4 \cdot 0{,}870 + 0{,}5 \cdot 0{,}130 = 0{,}413 = \widehat{t}_1,$$

pa se čini razumnim uzeti vrijednost $\widehat{t}_1 = 0{,}413$ kao procjenu nepoznatog parametra t, kada se kao vrijednost statistike $\overline{X}$ dobije $\frac{1}{5}$. Da se uzela u obzir samo apriorna razdioba, matematičko očekivanje iznosilo bi

$$E[T] = \overline{t} = 0{,}8 \cdot 0{,}4 + 0{,}2 \cdot 0{,}5 = 0{,}42.$$

Na temelju tabl. 4. općenito se može izračunati uvjetno očekivanje s.v. T uz uvjet da je statistika $\overline{X}$ poprimila vrijednost $\frac{k}{5}$, pa je

$$(94) \qquad \widehat{t}_k = E\left[T/\overline{X} = \frac{k}{5}\right] = \sum_{i=1}^{2} t_i q_{i/k}, \quad k = 0, 1, 2, 3, 4, 5,$$

i $\widehat{t}_k$ se uzima kao procjena za nepoznati parametar t u smislu Bayesove metode. Izvedu li se računske operacije naznačene formulom (94), dobiva se tabl. 5.

Tablica 5.

k	0	1	2	3	4	5
$\widehat{t}_k$	0,409	0,413	0,418	0,425	0,433	0,443

Ako se, na primjer, opažanjem dobije relativna frekvencija $\frac{3}{5}$ $(k = 3)$, onda će se kao procjena za parametar t, u smislu Bayesove metode, uzeti vrijednost $\widehat{t}_3 = 0{,}425$, dok bi procjena nepoznatog parametra samo na temelju vrijednosti statistike $\overline{X}$, bez uzimanja u obzir apriorne razdiobe parametra, iznosila $\widehat{t} = \frac{3}{5} = 0{,}6$.

Ako bi se, umjesto diskretne razdiobe definirane tablicom 2, za nepoznati parametar t pretpostavila uniformna razdioba $U(0,1)$, kao apriorna razdioba, onda ulogu tabl. 2. preuzima odgovarajuća funkcija gustoće vjerojatnosti

$$(95) \qquad f_1(t) = \begin{cases} 0, & \text{za} \quad t \le 0 \text{ i } t \ge 1 \\ 1, & \text{za} \quad 0 < t < 1, \end{cases}$$

a umjesto formule (91) imamo

$$f\left(\frac{k}{5}, t\right) = f_1(t) \mathrm{P}\left(\overline{X} = \frac{k}{5} / T = t\right),$$

pri čemu $f\left(\frac{k}{5}, t\right)$ označuje gustoću vjerojatnosti slučajnog vektora $(\overline{X}, T)$. Budući da s.v. $5\overline{X} \sim \mathrm{B}(5, t)$, onda je

$$\mathrm{P}\left(\overline{X} = \frac{k}{5} / T = t\right) = \binom{5}{k} t^k (1-t)^{5-k}, \quad k = 0,1,2,3,4,5,$$

tako da se konačno dobiva

$$(96) \qquad f\left(\frac{k}{5}, t\right) = \begin{cases} \binom{5}{k} t^k (1-t)^{5-k} & , \text{ za } 0 < t < 1, \ k = 0,1,2,3,4,5 \\ 0 & , \text{ inače.} \end{cases}$$

To omogućuje da se dobije formula, analogna formuli (92), koja glasi

$$(97) \qquad \mathrm{P}\left(\overline{X} = \frac{k}{5}\right) = \binom{5}{k} \int_0^1 t^k (1-t)^{5-k} \, \mathrm{d}t = \frac{1}{6}, \quad k = 0,1,2,3,4,5,$$

dok je analogon formule (93)

$$(98) \qquad g_k(t) = \frac{f\left(\frac{k}{5}, t\right)}{\mathrm{P}\left(\overline{X} = \frac{k}{5}\right)} = 6 \binom{5}{k} t^k (1-t)^{5-k}, \quad 0 < t < 1, \ k = 0,1,2,3,4,5.$$

Sada se može izračunati i

$$(99) \qquad \mathrm{E}\left[T / \overline{X} = \frac{k}{5}\right] = \widehat{t}_k = \int_0^1 t g_k(t) \, \mathrm{d}t = 6 \binom{5}{k} \int_0^1 t^{k+1} (1-t)^{5-k} \, \mathrm{d}t,$$

tj. uvjetno očekivanje slučajne varijable T uz uvjet da je statistika $\overline{X}$ poprimila vrijednost $\frac{k}{5}$. Broj $\widehat{t}_k$ $(k = 0,1,2,3,4,5)$ uzima se kao procjena za nepoznati parametar t u smislu Bayesove metode.

Ako se, prema formuli (99), izračunaju odgovarajuće vrijednosti, dobiva se tabl. 6. U trećem retku tabl. 6. navedene su vrijednosti $\overline{x} = \frac{k}{5}$ statistike $\overline{X}$.

Tablica 6.

k	0	1	2	3	4	5
$\widehat{t}_k$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{3}{7}$	$\frac{4}{7}$	$\frac{5}{7}$	$\frac{6}{7}$
$\overline{x}$	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1

Ako se nepoznati parametar $t = p$ (proporcija neispravnih proizvoda) procjenjuje samo na temelju ML-procjenitelja $\overline{X}$ i ako se opažanjem dobije konkretna vrijednost $\overline{x} = \frac{2}{5} = 0{,}4$, onda bi se ta vrijednost uzela kao procjena za nepoznati parametar p. Ako se, pak, procjenjuje Bayesovom metodom, uz pretpostavku da parametru p, kao slučajnoj varijabli, pripada uniformna razdioba $U(0,1)$ i ako se dobije $\frac{2}{5}$ kao vrijednost statistike $\overline{X}$, onda će se kao procjena za nepoznati parametar p uzeti vrijednost $\widehat{t}_2 = \frac{3}{7} \approx 0{,}482$.

Zaključimo: Ako se nepoznati parametar p procjenjuje samo na temelju apriorne vjerojatnosne razdiobe, onda se čini najprirodnijim uzeti kao procjenu očekivanje te razdiobe, tj. vrijednost $\mathrm{E}[T] = 0{,}5$, jer je tada, prema (69) iz IV.6, očekivana kvadratna greška najmanja.

Ako se, pak, procjenjuje pomoću ML-procjenitelja $\overline{X}$, a da se ne uzima u obzir apriorna razdioba, onda su vrijednosti procjena navedene u trećem retku tabl. 6.

Ako se, pak, uzmu u obzir i apriorna vjerojatnosna razdioba nepoznatog parametra p i vrijednosti statistike $\overline{X}$, onda su odgovarajuće vrijednosti procjene parametra p navedene u drugom retku tabl. 6. i za njih se kaže da su određene u smislu Bayesove metode.

Prijeđimo sada na općenito razmatranje Bayesove metode procjene nepoznatog parametra $t \in \Theta$ u teorijskom modelu s klasom $\mathcal{P} = \{P_t : t \in \Theta\}$ dopuštenih vjerojatnosnih razdioba. Ovdje se pojavljuje još i apriorna vjerojatnosna razdioba Π na skupu Θ, tako da će se govoriti o slučajnoj varijabli T kojoj pripada vjerojatnosna razdioba Π.

Neka je na slučajnom uzorku $(X_1, \ldots, X_n)$ definirana statistika $Y = h(X_1, \ldots, X_n)$, pa se može promatrati slučajni vektor (T, Y), kojemu, dakako, pripada određena dvodimenzionalna razdioba vjerojatnosti. Tada se može govoriti i o uvjetnoj razdiobi slučajne varijable T uz uvjet da je s.v. Y poprimila vrijednost y, te o odgovarajućem uvjetnom očekivanju $\mathrm{E}[T/Y = y]$.

Budući da je $y \mapsto \mathrm{E}[T/Y = y] = \phi(y)$ funkcija od y, onda je $\widehat{T}_Y = \phi(Y) = \mathrm{E}[T/Y]$ slučajna varijabla ovisna o slučajnom uzorku $(X_1, \ldots, X_n)$, tj. određena statistika koja se zove *procjenitelj nepoznatog parametra t u smislu Bayesove metode*. Vrijednost te slučajne varijable

(100) $$\widehat{t}_y = \phi(y) = \mathrm{E}[T/Y = y]$$

uzima se kao *procjena za nepoznati parametar t u smislu Bayesove metode*. Broj $\widehat{t}_y$ označuje uvjetno očekivanje s.v. T uz uvjet da je na n-članom slučajnom uzorku dobivena vrijednost $y = h(x_1, \ldots, x_n)$ statistike Y.

13. primjer

Uzmimo da je $\mathcal{P} = \{\mathrm{N}(t, \sigma^2) : t \in \mathbf{R}, \}$, pri čemu se pretpostavlja da je σ^2 poznato. Treba procijeniti nepoznato očekivanje $\mu = t$ normalne razdiobe poznate varijance, uz dodatnu informaciju da je $\Pi = \mathrm{N}(\mu_0, \sigma_0^2)$ apriorna vjerojatnosna razdioba nepoznatog parametra t. Prirodno je da se kao statistika za procjenu nepoznatog parametra t uzme uzoračka aritmetička sredina $\overline{X}$, tj. da se stavi $Y = \overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Uvedimo sljedeće oznake:

$f(t,y)$ – gustoća vjerojatnosti slučajnog vektora (T, Y),
$f_1(t)$ – gustoća vjerojatnosti s.v. $T \sim \mathrm{N}(\mu_0, \sigma_0^2)$,
$f_2(y)$ – gustoća vjerojatnosti s.v. Y,
$p_y(t)$ – uvjetna gustoća vjerojatnosti s.v. T uz uvjet da je s.v. Y poprimila vrijednost y,
$q_t(y)$ – uvjetna gustoća vjerojatnosti s.v. Y uz uvjet da je s.v. T poprimila vrijednost t.

Tada je

(101) $$f_1(t) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu_0}{\sigma_0}\right)^2\right],$$

(102) $$q_t(y) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} \exp\left[-\frac{n}{2}\left(\frac{y-t}{\sigma}\right)^2\right].$$

Na temelju formule (28) iz V.3. može se pisati

$$f(t,y) = f_1(t) q_t(y),$$

dok se primjenom formule (22) iz V.3. dobiva

$$f_2(y) = \int_{-\infty}^{\infty} f(t,y)\, \mathrm{d}t = \int_{-\infty}^{\infty} f_1(t)\, q_t(y)\, \mathrm{d}t,$$

pa formula (26) iz V.3. omogućuje da se dobije

(103) $$p_y(t) = \frac{f(t,y)}{f_2(y)} = \frac{f_1(t)\, q_t(y)}{\int_{-\infty}^{\infty} f_1(t)\, q_t(y)\, \mathrm{d}t}.$$

Iz (101), (102) i (103), nakon nešto složenijeg računanja, dobiva se

(104) $$p_y(t) = \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu_2}{\sigma_2}\right)^2\right],$$

gdje je

$$\mu_2 = \frac{ny\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}, \quad \sigma_2^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2},$$

iz čega se razabire da je uvjetna razdioba s.v. T, uz pretpostavku da je statistika Y poprimila vrijednost y, normalna razdioba $\mathrm{N}(\mu_2, \sigma_2^2)$, a to znači da je

$$\phi(y) = \mathrm{E}[T/Y = y] = \mu_2.$$

Reći će se, prema tome, da je

(105) $$\widehat{T}_Y = \phi(Y) = \frac{nY\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2}$$

procjenitelj za nepoznato očekivanje $t = \mu$ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$ u smislu Bayesove metode, kada je apriorna razdioba parametra t normalna razdioba $\mathrm{N}(\mu_0, \sigma_0^2)$.

Imajući na umu da je $y = \overline{x}$ vrijednost uzoračke aritmetičke sredine, zanimljivo je primijetiti da se nepoznati parametar μ procjenjuje, u smislu Bayesove metode, vrijednošću

(106) $$\widehat{t}_y = \frac{n\overline{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}}\overline{x} + \frac{\frac{\sigma^2}{n}}{\sigma_0^2 + \frac{\sigma^2}{n}}\mu_0,$$

koja se može shvatiti kao ponderirani zbroj vrijednosti $\overline{x}$ ML-procjenitelja $\overline{X}$ parametra μ i vrijednosti $\mu_0 = \mathrm{E}[T]$ (očekivanje apriorne razdiobe). Iz (106) se vidi da se s porastom veličine uzorka n "težina" $\frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}}$ ML-procjene sve više povećava i za $n \to \infty$ postaje jedan, dok se "težina" $\frac{\frac{\sigma^2}{n}}{\sigma_0^2 + \frac{\sigma^2}{n}}$ apriorne procjene sve više smanjuje i za $n \to \infty$ postaje nula. Može se, dakle, zaključiti da je pri velikom broju mjerenja ($n \to \infty$) procjena u smislu Bayesove metode praktički jednaka ML-procjeni nepoznatog parametra μ.

Zadaci

1. Neka funkcija gubitka ima opći oblik $L(\widehat{t}, t) = \max_{x \in \mathbf{R}} |F_{\widehat{t}}(x) - F_t(x)|$, gdje je F f.r.v.

a) Uzmu li se eksponencijalne razdiobe $\mathrm{Ex}(t)$ $(t > 0)$ kao klasa dopuštenih razdioba, onda je $L(\widehat{t}, t) = \dfrac{|\widehat{t} - t|}{t}$. Dokažite!

b) Uzmu li se uniformne razdiobe $\mathrm{U}(0, t)$ $(t > 0)$ kao klasa dopuštenih razdioba, onda je

$$L(\widehat{t}, t) = \begin{cases} 1 - \dfrac{\widehat{t}}{t}, & \text{za} \quad \widehat{t} \le t \\ 1 - \dfrac{t}{\widehat{t}}, & \text{za} \quad \widehat{t} > t. \end{cases}$$

Dokažite!

2. Dokažite da nepristrani procjenitelj $\widehat{T}$, koji ima minimalnu varijancu, zadovoljava minimaks-princip izražen formulom (24).

3. Neka je $\overline{X}$ uzoračka aritmetička sredina n-članoga slučajnog uzorka za slučajnu varijablu X, čiji su μ_k centralni momenti. Dokažite da za centralne momente $\overline{\mu}_k$ slučajne varijable $\overline{X}$ vrijede formule: a) $\overline{\mu}_1 = 0$, b) $\overline{\mu}_2 = \dfrac{1}{n}\mu_2$, c) $\overline{\mu}_3 = \dfrac{1}{n^2}\mu_3$, d) $\overline{\mu}_4 = \dfrac{1}{n^3}[\mu_4 + 3(n-1)\mu_2^2]$.

4. Izvedite formule (30) i (31).

5. Dokažite da je $\mathrm{Cov}(\overline{X}, \widehat{\Sigma}^2) = \dfrac{n-1}{n^2}\mu_3$, gdje je μ_3 treći centralni moment slučajne varijable X.

6. Ako slučajnoj varijabli X pripada simetrična razdioba, onda su $\overline{X}$ i $\widehat{\Sigma}^2$ nekorelirane slučajne varijable. Dokažite!

7. Neka je $\widehat{T} = a\widehat{\Sigma}^2$ $(a > 0)$ procjenitelj za nepoznatu varijancu $\sigma^2 = t$ normalne razdiobe $\mathrm{N}(\mu, t)$. Dokažite da:

a) $R(t) = \mathrm{E}[(\widehat{T} - t)^2] = \dfrac{1}{n^2}[(n^2 - 1)a^2 - 2n(n-1)a + n^2]t$, $t > 0$,

b) $\min_{a > 0} R(t) = \dfrac{2}{n+1}t$, postiže se za $a = \dfrac{n}{n+1}$,

c) u klasi procjenitelja $\mathcal{T} = \{\widehat{T} = a\widehat{\Sigma}^2 : a > 0\}$ ne postoji procjenitelj koji zadovoljava minimaks-princip.

8. Dokažite da je uzorački ishodišni moment $\widehat{A}_r = \dfrac{1}{n}\sum_{i=1}^{n} X_i^r$ $(r = 1, 2, \ldots)$ nepristrani procjenitelj za teorijski ishodišni moment β_r i da vrijedi $\mathrm{V}[\widehat{A}_r] = \dfrac{1}{n}(\beta_{2r} - \beta_r^2)$.

9. Dokažite da za procjenitelj $\widehat{\Lambda}$ za parametar λ Poissonove razdiobe $\mathrm{Po}(\lambda)$, definiran formulom (39), vrijedi $\mathrm{E}[\widehat{\Lambda}] = \lambda$ i $\mathrm{V}[\widehat{\Lambda}] = \dfrac{1}{n}\lambda$.

10. Nađite ML-procjenitelj za parametar p $(0 < p < 1)$ geometrijske razdiobe (v. IV.3).

11. Dokažite da su $\widehat{A} = \min(X_1, \ldots, X_n)$ i $\widehat{B} = \max(X_1, \ldots, X_n)$ ML-procjenitelji za parametre a i b uniformne razdiobe $\mathrm{U}(a, b)$ $(a < b)$.

12. Nađite ML-procjenitelj za parametar p $(0 < p < 1)$ binomne razdiobe $\mathrm{B}(m, p)$ uz pretpostavku da je m poznato.

13. Nađite ML-procjenitelj za parametar t $(t \in \mathbf{R})$ kontinuirane razdiobe vjerojatnosti čija f.g.v. glasi

$$f_t(x) = \begin{cases} 0, & \text{za} \quad x \le t \\ \exp[-(x - t)] & \text{za} \quad x > t. \end{cases}$$

14. Nađite ML-procjenitelj za vektorski parametar $\boldsymbol{t} = (t_1, t_2)$ kontinuirane razdiobe vjerojatnosti čija f.g.v. glasi

$$f_{\boldsymbol{t}}(x) = \begin{cases} 0, & \text{za} \quad x < t_1 \\ \dfrac{1}{t_2}\exp\left(-\dfrac{x - t_1}{t_2}\right), & \text{za} \quad x \ge t_1, \end{cases}$$

gdje je $\boldsymbol{\Theta} = \{(t_1, t_2) \in \mathbf{R}^2 : t_1 \in \mathbf{R}, t_2 > 0\}$.

15. Nađite ML-procjenitelj za vektorski parametar $\boldsymbol{t} = (t_1, t_2)$ vjerojatnosne razdiobe, čija f.r.v. glasi

$$F_{\boldsymbol{t}}(x) = \begin{cases} 0, & \text{za} \quad x < t_1 \\ 1 - \left(\dfrac{t_1}{x}\right)^{t_2} & \text{za} \quad x \ge t_1. \end{cases}$$

gdje je $\boldsymbol{\Theta} = \{(t_1, t_2) \in \mathbf{R}^2 : t_1 > 0, t_2 > 0\}$.

16. Izvedite formulu (46) imajući na umu da funkcija gubitka glasi

$$L(\widehat{t}, t) = (\widehat{\mu} - \mu)^2 + (\widehat{\sigma}^2 - \sigma^2)^2$$

i da je $R(t) = E[L(\widehat{\boldsymbol{T}}, t)]$, pri čemu je $\widehat{\boldsymbol{T}} = (\overline{X}, \widehat{\Sigma}^2)$.

17. Izvedite formulu (48).

18. Dokažite da je statistika $\overline{X}^2$ ML-procjenitelj za varijancu eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$.

19. Nađite funkciju rizika za procjenitelj iz zad. 18. i pokažite da je za velike n $(n \to \infty)$ ML-procjenitelj $\overline{X}^2$ približno četiri puta bolji procjenitelj od $\widehat{\Sigma}^2$.

20. Ako je $\widehat{T}$ ML-procjenitelj za parametar t i $t \mapsto h(t)$, $t \in \Theta$ strogo rastuća funkcija, tada je $h(\widehat{T})$ ML-procjenitelj za $h(t)$. Dokažite!

21. Neka su $\widehat{T}_1$ i $\widehat{T}_2$ nezavisni i nepristrani procjenitelji za parametar t, kojima pripadaju konačne varijance $\mathrm{V}[\widehat{T}_1]$ i $\mathrm{V}[\widehat{T}_2]$.

a) Dokažite da je $\widehat{T} = c_1\widehat{T}_1 + c_2\widehat{T}_2$ $(c_1, c_2 \in \mathbf{R}, c_1 + c_2 = 1)$ nepristrani procjenitelj za t.

b) Dokažite da varijanca $V[\widehat{T}]$ postaje minimalna ako vrijedi $c_1 V[\widehat{T}_1] = c_2 V[\widehat{T}_2]$.

22. Dokažite formule (75) i (76).

23. Dokažite Rao-Cramerovu nejednakost za slučaj diskretnih razdioba.

24. Odredite efikasnost uzoračke aritmetičke sredine $\overline{X}$ kao procjenitelja za parametar:

a) μ normalne razdiobe $N(\mu, \sigma^2)$,

b) p binomne razdiobe $B(1, p)$,

c) $a = \frac{1}{2}t$ uniformne razdiobe $U(0, t)$.

25. Odredite efikasnost statistike $\widehat{T}_n = \frac{n+1}{n}\max(X_1, \ldots, X_n)$ kao procjenitelja za parametar t uniformne razdiobe $U(0, t)$.

26. Nađite Rao-Cramerovu donju granicu za varijance procjenitelja za parametar t ($t \in \mathbf{R}$) vjerojatnosne razdiobe čija f.g.v. glasi

$$f_t(x) = \frac{1}{\pi[1 + (x - t)^2]}, \quad x \in \mathbf{R}.$$

27. Primjenom CGT-a dokažite da se za velike m binomna razdioba $B(m, p)$ može aproksimirati normalnom razdiobom $N(mp, mp(1 - p))$.

28. Primjenom CGT-a nađite vjerojatnost da zbroj od $n = 12$ brojeva uzetih iz intervala $\langle 0, 1 \rangle$ u skladu s uniformnom razdiobom $U(0, 1)$ bude veći od 10.

29. Neka je $\mathcal{P} = \{B(1, p) : 0 < p < 1\}$ klasa dopuštenih razdioba vjerojatnosti i neka je apriorna razdioba parametra $t = p$ beta-razdioba sa zadanim parametrima α i β. Primjenjujući statistiku $Y = n\overline{X} = X_1 + \ldots + X_n$ i kvadratnu grešku kao funkciju gubitka, nađite Bayesov procjenitelj za nepoznati parametar t binomne razdiobe $B(1, t)$.

30. Neka je $\mathcal{P} = \{Po(\lambda) : \lambda > 0\}$ klasa dopuštenih razdioba vjerojatnosti i neka je apriorna razdioba parametra $t = \lambda$ eksponencijalna razdioba $Ex(\alpha)$. Primjenjujući statistiku $Y = n\overline{X}$ i kvadratnu grešku kao funkciju gubitka, nađite Bayesov procjenitelj za parametar λ.

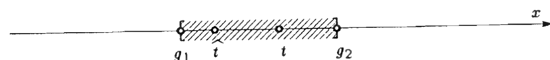
Pregled najvažnijih procjenitelja

Pretpostavljena klasa razdioba	Parametar	Procjenitelj	Svojstva
$B(1, p)$	p	$\overline{X}$	ML-procjenitelj, nepristran, konzistentan, najefikasniji, asimptotski normalan
$Po(\lambda)$	λ	$\overline{X}$	ML-procjenitelj, nepristran, konzistentan, najefikasniji, asimptotski normalan
$Ex(\alpha)$	$\alpha_0 = \frac{1}{\alpha}$	$\overline{X}$	ML-procjenitelj, nepristran, konzistentan, najefikasniji, asimptotski normalan
$U(0, t)$	t	$2\overline{X}$	nepristran, konzistentan, asimptotski normalan
		$\max(X_1, \ldots, X_n)$	ML-procjenitelj, konzistentan
$N(\mu, \sigma^2)$	μ	$\overline{X}$	ML-procjenitelj, nepristran, konzistentan, najefikasniji, normalan
	σ^2	$\widehat{\Sigma}^2$	ML-procjenitelj, konzistentan
		S^2	nepristran, konzistentan, asimptotski efikasan, asimptotski normalan
postoji konačna varijanca	μ	$\overline{X}$	nepristran, konzistentan asimptotski normalan
postoji konačni četvrti centralni moment	σ^2	S^2	nepristran, konzistentan asimptotski normalan
postoji konačni $2r$-ti centralni moment	β_r	$\frac{1}{n}\sum_{i=1}^{n} X_i^r$	nepristran, konzistentan asimptotski normalan

VII. Intervali povjerenja

1. Uvod u problematiku

U VI. poglavlju razmatran je problem procjene nepoznatog parametra vjerojatnosne razdiobe sa svrhom da se definira dobar procjenitelj čija vrijednost služi kao aproksimacija nepoznatog parametra. Kaže se još da je time izvršena ***točkasta procjena*** nepoznatog parametra, za razliku od ***intervalne procjene*** o kojoj će biti riječ u ovom poglavlju. Naime, na temelju niza podataka $x_1, \dots, x_n$, treba odrediti interval $\langle g_1, g_2 \rangle$ koji će imati svojstvo da s velikom vjerojatnošću "pokriva" nepoznati parametar t.



Slika 6. Skica intervalne procjene parametra

Uzme li se $\hat{t} \in \langle g_1, g_2 \rangle$ i stavi $t \approx \hat{t}$, onda se s velikom vjerojatnošću može jamčiti da apsolutna greška pri aproksimaciji nepoznatog parametra t vrijednošću $\hat{t}$ nije veća od $\delta = |g_2 - g_1|$. Rubovi g_1 i g_2 intervala $\langle g_1, g_2 \rangle$ ovisit će, dakako, o izmjerenim podacima $x_1, \dots, x_n$, tako da se može pisati

$$(1) \qquad g_1 = h_1(x_1, \dots, x_n), \quad g_2 = h_2(x_1, \dots, x_n),$$

pri čemu su h_1 i h_2 određene funkcije n realnih varijabli takve da vrijedi

$$h_1(x_1, \dots, x_n) < h_2(x_1, \dots, x_n), \quad (x_1, \dots, x_n) \in \mathbf{R}^n.$$

No, kako je već rečeno u VI.1, niz $(x_1, \dots, x_n)$ može se shvatiti kao vrijednost slučajnog vektora $(X_1, \dots, X_n)$, gdje su $X_1, \dots, X_n$ nezavisne slučajne varijable sa zajedničkom vjerojatnosnom razdiobom $P_t \in \mathcal{P}$ ($\mathcal{P}$ je klasa dopuštenih razdioba vjerojatnosti), tako da se g_1 i g_2 mogu razmatrati kao vrijednosti statistika

$$(2) \qquad G_1 = h_1(X_1, \dots, X_n), \quad G_2 = h_2(X_1, \dots, X_n).$$

Statistike G_1 i G_2 su, dakako, određene slučajne varijable kojima, ovisno o parametru t, pripadaju odgovarajuće razdiobe vjerojatnosti, pa se može postaviti zahtjev

$$(3) \qquad \mathrm{P}(G_1 < t < G_2) \geq \gamma, \quad t \in \Theta,$$

gdje je γ $(0 < \gamma < 1)$ zadani realni broj.

Ako su rubovi g_1, g_2 intervala $\langle g_1, g_2 \rangle$ dobiveni kao vrijednosti statistika G_1 i G_2, pri čemu je zadovoljena relacija (3), onda se kaže da je $\langle g_1, g_2 \rangle$ ***interval povjerenja pouzdanosti*** γ za nepoznati parametar t.

Broj γ zove se još i *razina povjerenja* i obično se uzima $\gamma = 0{,}95$ ili $0{,}99$, pa se govori o 95 % ili 99 % pouzdanosti izvedene intervalne procjene, odnosno o 95-postotnoj ili 99-postotnoj razini povjerenja.

Ako se, na primjer, određuje interval povjerenja uz pouzdanost od 95 %, onda se može očekivati da će se, prilikom višestrukog uzimanja slučajnog uzorka, bar u 95 % slučajeva dobiti interval povjerenja koji će sadržavati nepoznati parametar, odnosno da će u najviše 5 % slučajeva nepoznati parametar ostati izvan dobivenog intervala povjerenja. Naravno da uz 99 % pouzdanosti imamo veću sigurnost da će interval povjerenja "pokriti" nepoznati parametar, međutim taj će interval biti, što će se vidjeti kasnije, širi od odgovarajućeg intervala povjerenja pouzdanosti od 95 %.

Pitanje izbora veličine pouzdanosti γ, tj. da li uzeti za γ vrijednost 0,95 ili 0,99 ili nešto treće i nije matematičko teorijsko pitanje, već je to stvar praktične prirode problema i procjene stručnjaka o utjecaju pojedinih faktora (pouzdanost, ekonomičnost, veličina greške i sl.) na konačnu odluku i njezine praktične posljedice.

Prema tome, teorijski gledano, problem određivanja intervala povjerenja zadane pouzdanosti γ za nepoznati parametar t koji postoji u danoj klasi dopuštenih vjerojatnosnih razdioba $\mathcal{P} = \{P_t : t \in \Theta\}$ sastoji se u definiranju statistika (2) koje ispunjavaju uvjet (3). Praktički se interval povjerenja određuje tako da se načini n nezavisnih mjerenja $x_1, \dots, x_n$ promatrane slučajne varijable X i izračunaju vrijednosti g_1 i g_2 statistika G_1 i G_2.

Primijetimo odmah da statistike G_1 i G_2 nisu jednoznačno određene uvjetom (3). Ako, naime, postoje statistike G_1 i G_2 koje zadovoljavaju uvjet

$$(4) \qquad \mathrm{P}(G_1 < t < G_2) = \gamma, \quad t \in \Theta,$$

onda one zadovoljavaju i uvjet (3), a očigledno je da mogu postojati statistike G_1 i G_2 koje ne ispunjavaju uvjet (4), a da zadovoljavaju (3).

Ako statistike G_1 i G_2 zadovoljavaju uvjet (4), onda se dobiva tzv. ***najuži interval povjerenja*** zadane pouzdanosti γ.

Ne bi se moglo reći da postoji neka univerzalna metoda za nalaženje intervala povjerenja, ali se mnogi konkretni problemi, kako će se vidjeti u nastavku, mogu rješavati na sljedeći način: Pretpostavimo da je $\hat{T}$ određeni procjenitelj za nepoznati parametar t, pa kako je $\hat{T}$ slučajna varijabla, kojoj pripada odgovarajuća razdioba vjerojatnosti, ovisna, dakako, o parametru t, može se postaviti zahtjev

$$(5) \qquad \mathrm{P}(c_1(t) < \hat{T} < c_2(t)) = \gamma, \quad t \in \Theta,$$

pri čemu su $t \mapsto c_1(t)$ i $t \mapsto c_2(t)$ monotone funkcije za koje vrijedi da je $c_1(t) \leq c_2(t)$ $(t \in \Theta)$, što osigurava postojanje odgovarajućih im inverznih funkcija h_1 i h_2, tako da se (5) može zapisati u obliku

$$(6) \qquad \mathrm{P}(h_1(\hat{T}) < t < h_2(\hat{T})) = \gamma.$$

Budući da su $h_1(\widehat{T})$ i $h_2(\widehat{T})$, kao funkcije slučajne varijable $\widehat{T}$, također određene slučajne varijable, uspoređivanjem (4) i (6) vidi se da je $G_1 = h_1(\widehat{T})$ i $G_2 = h_2(\widehat{T})$ i time je načelno riješeno pitanje rubova intervala povjerenja pouzdanosti γ za nepoznati parametar t. Problem se, prema tome, sveo na to da se odredi razdioba vjerojatnosti procjenitelja $\widehat{T}$ i da se nađu funkcije c_1 i c_2 koje ispunjavaju uvjet (5).

Taj postupak ilustrirat će se idućim primjerom.

1. primjer

Pretpostavlja se da je $\mathcal{P} = \{\mathrm{N}(t, \sigma^2) : t \in \mathbf{R}\}$, tj. treba odrediti interval povjerenja pouzdanosti γ za nepoznato očekivanje $\mu = t$ normalne razdiobe, čija je varijanca σ^2 poznata.

U praksi se takva situacija pojavljuje kada se, na primjer, promatra proces proizvodnje određenog proizvoda (vijka, pločice i sl.) ne nekom stroju koji je podešen tako da proizvod ima propisanu dimenziju X (duljinu, debljinu i sl.). Zbog brojnih različitih utjecaja dimenzija X je slučajna varijabla za koju se pretpostavlja da ima normalnu razdiobu $\mathrm{N}(\mu, \sigma^2)$, pri čemu μ ovisi o radnoj podešenosti stroja, dok σ^2 ovisi o preciznosti i tvorničkoj kvaliteti stroja, što je obično unaprijed poznato. Na temelju niza mjerenja $x_1, \ldots, x_n$ dimenzije X na izrađenim proizvodima, želi se provjeriti je li stroj ispravno podešen, tj. da li se parametar μ nalazi u granicama tolerancije, dakako, unaprijed zadanima.

Teorijski dio problema može se rješavati ovako: Polazi se od spoznaje da je $\widehat{T} = \overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ procjenitelj za nepoznati parametar $t = \mu$ i da $\overline{X} \sim \mathrm{N}\left(t, \frac{1}{n}\sigma^2\right)$ (v. VI.4. (52)). Treba još, u skladu sa (5), pronaći funkcije c_1 i c_2.

Budući da je (v. VI.5. (46))

(7) $$\mathrm{P}\left(t - \lambda\frac{\sigma}{\sqrt{n}} < \overline{X} < t + \lambda\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(\lambda) - 1, \quad \lambda > 0,$$

odmah se nameće ideja da se uzme

(8) $$c_1(t) = t - \lambda\frac{\sigma}{\sqrt{n}}, \quad c_2(t) = t + \lambda\frac{\sigma}{\sqrt{n}},$$

pa se vidi da su to strogo rastuće funkcije parametra $t \in \mathbf{R}$ te se (7) može pisati u obliku

(9) $$\mathrm{P}\left(\overline{X} - \lambda\frac{\sigma}{\sqrt{n}} < t < \overline{X} + \lambda\frac{\sigma}{\sqrt{n}}\right) = 2\Phi(\lambda) - 1.$$

Uspoređujući (9) i (6), najprije se vidi da se λ određuje iz zahtjeva $2\Phi(\lambda) - 1 = \gamma$, iz čega proizlazi

(10) $$\lambda = z_\gamma = \Phi^{-1}\left(\frac{1+\gamma}{2}\right).$$

Zatim se vidi da su tražene statistike G_1 i G_2, kao slučajni rubovi intervala povjerenja pouzdanosti γ, dane formulama

(11) $$G_1 = \overline{X} - z_\gamma\frac{\sigma}{\sqrt{n}}, \quad G_2 = \overline{X} + z_\gamma\frac{\sigma}{\sqrt{n}}.$$

Zanimljivo je uočiti da su G_1 i G_2, osim što su funkcije procjenitelja $\overline{X}$, također ovisne o pouzdanosti γ i veličini uzorka n, dok širina intervala povjerenja

(12) $$\delta = |G_2 - G_1| = 2z_\gamma\frac{\sigma}{\sqrt{n}}$$

ovisi samo o γ i n i nije slučajna varijabla.

Kao što je poznato, funkcija Φ i njoj inverzna funkcija Φ^{-1} odnose se na standardnu normalnu razdiobu $\mathrm{N}(0,1)$, tako da se proračun veličine z_γ može izvesti primjenom tabl. III. u Dodatku. Za najčešće vrijednosti pouzdanosti γ, odgovarajuće vrijednosti z_γ, izračunane prema formuli (10), navedene su u tabl. 1.

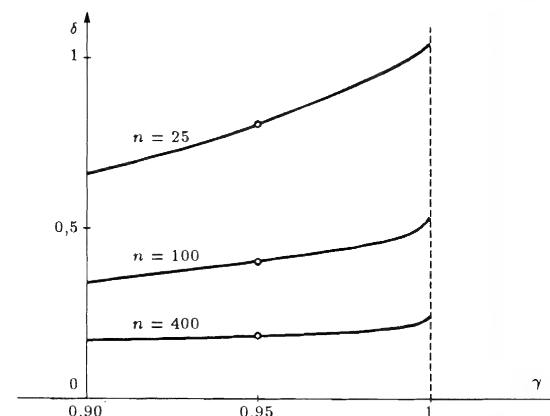
Tablica 1.

γ	0,90	0,95	0,99
z_γ	1,65	1,96	2,58

Tablica 2.

n \ γ	0,90	0,95	0,99
25	0,66	0,78	1,03
100	0,33	0,39	0,52
400	0,16	0,20	0,26

Na temelju (12) i tabl. 1 može se uočiti ovisnost između širine intervala povjerenja, pripadne pouzdanosti i veličine uzorka, što je prikazano tablicom 2, u kojoj su za odabrane n i γ navedene vrijednosti veličine $\delta = \frac{2}{\sqrt{n}}z_\gamma$ koje pokazuju širinu intervala povjerenja za $\sigma = 1$. Odgovarajući grafički prikaz dan je na sl. 7.



Slika 7. Skica ovisnosti veličina n, γ i δ

Sada se može navesti i primjer za konkretno određivanje intervala povjerenja. Recimo da je $\sigma = 2$ i da je na uzorku veličine $n = 25$ dobivena vrijednost uzoračke aritmetičke sredine $\overline{x} = 12$, pa će tada, prema (11) i tabl. 1, interval povjerenja

pouzdanosti $\gamma = 0{,}95$ imati granice

$$g_1 = \overline{x} - z_\gamma \frac{\sigma}{\sqrt{n}} = 12 - 1{,}96 \cdot \frac{2}{5} = 11{,}22,$$
$$g_2 = \overline{x} + z_\gamma \frac{\sigma}{\sqrt{n}} = 12 + 1{,}96 \cdot \frac{2}{5} = 12{,}78.$$

Zaključak je, dakle, da se nepoznato očekivanje μ nalazi u intervalu $\langle 11{,}22\ ;\ 12{,}78\rangle$, ali taj zaključak nije apsolutno siguran, već ima pouzdanost od 95 %, što znači da primjenom opisanog postupka možemo očekivati 5 % pogrešnih zaključaka. Ako bi se toleriralo samo 1 % pogrešnih zaključaka, tj. ako se uzme $\gamma = 0{,}99$, onda je $z_\gamma = 2{,}58$ i za rubove intervala povjerenja dobivaju se vrijednosti

$$g_1 = 12 - 2{,}58 \cdot \frac{2}{5} = 10{,}97, \qquad g_2 = 12 + 2{,}58 \cdot \frac{2}{5} = 13{,}03.$$

Formula (12) omogućuje i da se, za unaprijed određenu dopuštenu grešku δ_0 pri aproksimaciji nepoznatog parametra t vrijednošću $\hat{t}$ iz intervala povjerenja pouzdanosti γ, odredi potrebna veličina n slučajnog uzorka. Stavi li se, naime, $\delta = \delta_0$, iz (12) se dobiva

(13) $$n = 4z_\gamma^2 \left(\frac{\sigma}{\delta_0}\right)^2.$$

Ako se kao jedinica mjere za grešku uzme standardna devijacija σ, tj. stavi li se $\delta_0 = k\sigma$ $(k \geq 0)$, (13) postaje

(14) $$n = \frac{4z_\gamma^2}{k^2},$$

pa se vidi da je broj n potrebnih mjerenja za određivanje intervala povjerenja pouzdanosti γ obrnuto proporcionalan s kvadratom dopuštene greške. Tako, na primjer, ako se uz pouzdanost $\gamma = 0{,}95$ tolerira greška od polovine standardne devijacije, tj. ako je $k = 0{,}5$, onda se iz tabl. 1. i formule (14) zaključuje da je potrebno načiniti bar $n = 62$ mjerenja promatrane slučajne varijable X.

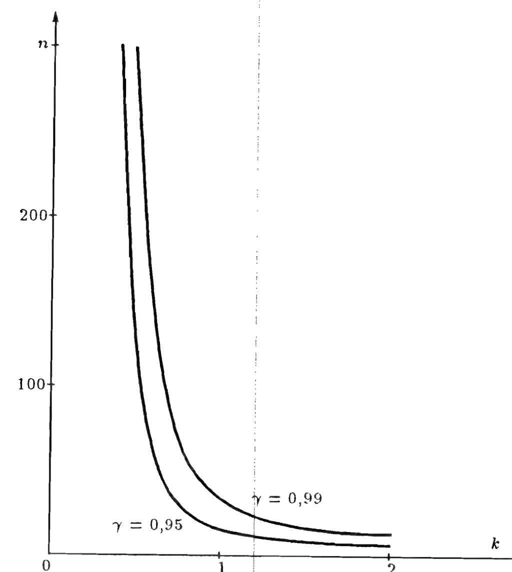
Ovisnost širine intervala povjerenja, odnosno dopuštene greške δ_0 i veličine uzorka n zorno je prikazana na sl. 8.

2. primjer

U vezi sa 1. primjerom iz VI.1. može se postaviti zadatak da se odredi interval povjerenja pouzdanosti γ za nepoznati parametar t uniformne razdiobe $\mathrm{U}(0,t)$ $(t > 0)$. Pokazano je da ML-procjenitelju $\widehat{T}_2 = \max(X_1, \ldots, X_n)$ za nepoznati parametar t pripada vjerojatnosna razdioba, čija je f.r.v. izražena formulom (19) iz VI.1. Primjenom relacije (5) na procjenitelj $\widehat{T}_2$ dobiva se

$$P(c_1(t) < \widehat{T}_2 < c_2(t)) = \left[\frac{c_2(t)}{t}\right]^n - \left[\frac{c_1(t)}{t}\right]^n = \gamma,$$

pa je riječ o tome da se nađu monotone funkcije c_1 i c_2 koje zadovoljavaju gornju jednadžbu. Budući da je u uniformnoj razdiobi $\mathrm{U}(0,t)$ svaki izmjereni podatak



Slika 8. Skica ovisnosti greške (k) i veličine uzorka (n)

$x_i \leq t$ $(i = 1, \ldots, n)$, bit će i $\hat{t}_2 = \max(x_1, \ldots, x_n) \leq t$, pa je prikladno uzeti $c_2(t) = t$ i $c_1(t) = kt$, te zahtijevati da se k $(0 < k < 1)$ odredi tako da vrijedi

$$\left[\frac{c_2(t)}{t}\right]^n - \left[\frac{c_1(t)}{t}\right]^n = 1 - k^n = \gamma,$$

iz čega odmah proizlazi da je $k = \sqrt[n]{1-\gamma}$. Stoga se može pisati

$$\mathrm{P}(t\sqrt[n]{1-\gamma} < \widehat{T}_2 < t) = \gamma,$$

odnosno

$$\mathrm{P}\left(\widehat{T}_2 < t < \frac{\widehat{T}_2}{\sqrt[n]{1-\gamma}}\right) = \gamma.$$

Uspoređivanjem te relacije s relacijom (4) vidi se da su slučajni rubovi intervala povjerenja pouzdanosti γ za parametar t uniformne razdiobe $\mathrm{U}(0,t)$ dani formulama

$$G_1 = \widehat{T}_2, \quad G_2 = \frac{\widehat{T}_2}{\sqrt[n]{1-\gamma}}.$$

Ako se, na primjer, uzme $\gamma = 0{,}95$ i $n = 25$, onda je $\sqrt[n]{1-\gamma} = \sqrt[25]{0{,}05} \approx 0{,}89$, pa ako je $\hat{t}_2$ maksimalna vrijednost u nizu mjerenja $x_1, \ldots, x_{25}$, onda odgovarajući rubovi intervala povjerenja za parametar t jesu

$$g_1 = \hat{t}_2, \quad g_2 = \frac{\hat{t}_2}{0{,}89} = 1{,}12\hat{t}_2.$$

U razmotrenom je primjeru širina intervala povjerenja

$$\Delta = |G_2 - G_1| = \hat{T}_2 \left(\frac{1}{\sqrt[n]{1-\gamma}} - 1 \right),$$

pa se vidi da je vrijednost širine intervala povjerenja

$$\delta = \hat{t}_2 \left(\frac{1}{\sqrt[n]{1-\gamma}} - 1 \right)$$

proporcionalna vrijednosti $\hat{t}_2$ procjenitelja $\hat{T}_2$ s faktorom proporcionalnosti

$$d = \frac{1}{\sqrt[n]{1-\gamma}} - 1.$$

Konkretniji pregled odnosa veličina n, d i γ može se dobiti u tabl. 3.

Tablica 3.

n	d		
	$\gamma = 0,90$	$\gamma = 0,95$	$\gamma = 0,99$
2	2,16	3,47	9
5	0,58	0,82	1,51
10	0,26	0,35	0,58
20	0,12	0,16	0,26
50	0,05	0,06	0,10
100	0,02	0,03	0,05

2. Intervali povjerenja za parametre normalne razdiobe

U 1. primjeru riješen je problem određivanja intervala povjerenja za parametar μ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$ uz pretpostavku da je parametar σ^2 poznat. Sada će se, međutim, postaviti zadatak da se nađu postupci za određivanje intervala povjerenja zadane pouzdanosti γ za oba nepoznata parametra μ i σ^2 na temelju niza mjerenja $x_1, \ldots, x_n$ slučajne varijable $X \sim \mathrm{N}(\mu, \sigma^2)$.

Budući da je ovdje riječ o dva parametra μ i σ^2, stavit će se $\boldsymbol{t} = (\mu, \sigma^2)$ i govoriti da je $\boldsymbol{t}$ vektorski parametar, pa se najprije mora reći što će se općenito razumijevati pod intervalom povjerenja pouzdanosti γ za nepoznati vektorski parametar $\boldsymbol{t} = (t_1, \ldots, t_k) \in \boldsymbol{\Theta} \subseteq \mathbf{R}^k$ $(k \in \mathbf{N})$. Ako su, dakle, G_{j1} i G_{j2} $(j = 1, \ldots, k)$ statistike koje zadovoljavaju uvjet

(15) $$\mathrm{P}(G_{11} < t_1 < G_{12}, \ldots, G_{k1} < t_k < G_{k1}) \geq \gamma, \quad \boldsymbol{t} \in \boldsymbol{\Theta},$$

a g_{j1} i g_{j2} $(j = 1, \ldots, k)$ su vrijednosti odgovarajućih statistika na izmjerenom nizu podataka $x_1, \ldots, x_n$, onda se skup

(16) $$\boldsymbol{I}_k = \langle g_{11}, g_{12} \rangle \times \ldots \times \langle g_{k1}, g_{k2} \rangle \subseteq \mathbf{R}^k,$$

dobiven, dakle, kao Kartezijev produkt intervala $\langle g_{11}, g_{12} \rangle, \ldots, \langle g_{k1}, g_{k2} \rangle$, zove *simultani interval povjerenja* pouzdanosti γ $(0 < \gamma < 1)$ za nepoznati vektorski parametar $\boldsymbol{t}$.

Interpretacija simultanog intervala povjerenja pouzdanosti γ za nepoznati vektorski parametar $\boldsymbol{t}$ slična je onoj za obični interval povjerenja. Ako je, recimo, $\gamma = 0{,}95$, onda se može očekivati da će se, prilikom višekratnog ponavljanja slučajnog eksperimenta koji se sastoji od uzimanja n-članog uzorka i određivanja odgovarajućeg skupa $\boldsymbol{I}_k \subseteq \mathbf{R}^k$ definiranog u (16), bar u 95 % slučajeva dobiti takav $\boldsymbol{I}_k$ koji će pokriti točku $\boldsymbol{t} \in \mathbf{R}^k$, odnosno da će u najviše 5 % slučajeva točka $\boldsymbol{t}$, tj. nepoznati vektorski parametar $\boldsymbol{t} = (t_1, \ldots, t_k)$, ostati izvan skupa $\boldsymbol{I}_k$.

Tako općenito formuliran problem vrlo je teško riješiti i stoga se postavlja lakši zadatak da se odrede intervali povjerenja za svaku komponentu posebno, tj. da se nađu statistike G_{j1} i G_{j2} $(j = 1, \ldots, k)$ koje zadovoljavaju uvjete

(17) $$\mathrm{P}(G_{j1} < t_j < G_{j2}) \geq \gamma_j, \quad \boldsymbol{t} \in \boldsymbol{\Theta}.$$

Ako su g_{j1} i g_{j2} vrijednosti statistika G_{j1} i G_{j2} na izmjerenom nizu podataka $x_1, \ldots, x_n$, onda je $\langle g_{j1}, g_{j2} \rangle$ interval povjerenja pouzdanosti γ_j $(0 < \gamma_j < 1)$ parametra t_j.

Na temelju elementarnih svojstava vjerojatnosti slijedi da je

(18) $$\begin{aligned} &\mathrm{P}(G_{11} < t_1 < G_{12}, \ldots, G_{k1} < t_k < G_{k2}) \geq \\ &\geq 1 - [1 - \mathrm{P}(G_{11} < t_1 < G_{12}) + \ldots + 1 - \mathrm{P}(G_{k1} < t_k < G_{k2})], \end{aligned}$$

pa iz (15), (17) i (18) proizlazi

(19) $$\mathrm{P}(G_{11} < t_1 < G_{12}, \ldots, G_{k1} < t_k < G_{k2}) \geq 1 - [(1 - \gamma_1) + \ldots + (1 - \gamma_k)].$$

Relacija (19) pokazuje da se, na temelju pojedinačno određenih intervala povjerenja pouzdanosti γ_j za komponente t_j $(j = 1, \ldots, k)$, može dobiti simultani interval povjerenja pouzdanosti $\gamma = 1 - [(1 - \gamma_1) + \ldots + (1 - \gamma_k)]$ za vektorski parametar $\boldsymbol{t} = (t_1, \ldots, t_k)$.

Posebno, ako je $k = 2$ i $\gamma_1 = \gamma_2 = 0{,}95$, te ako su $\langle g_{11}, g_{12} \rangle$ i $\langle g_{21}, g_{22} \rangle$ intervali povjerenja komponenata t_1 i t_2 pouzdanosti 95 %, onda je pravokutnik $\boldsymbol{I}_2 = \langle g_{11}, g_{12} \rangle \times \langle g_{21}, g_{22} \rangle$ simultani interval povjerenja pouzdanosti $\gamma = 1 - 0{,}05 - 0{,}05 = 0{,}90$ za vektorski parametar $\boldsymbol{t} = (t_1, t_2)$.

Sada se možemo vratiti zadatku pojedinačnog određivanja intervala povjerenja za parametre normalne razdiobe $t_1 = \mu$ i $t_2 = \sigma^2$, uz pretpostavku da je $\boldsymbol{t} = (\mu, \sigma^2) \in \boldsymbol{\Theta}$ nepoznati vektorski parametar, pri čemu je $\boldsymbol{\Theta} = \{(\mu, \sigma^2) \in \mathbf{R}^2 : \mu \in \mathbf{R}, \sigma^2 > 0\}$.

Odmah se može primijetiti da se za određivanje intervala povjerenja za parametar μ neće moći, kao u 1. primjeru, neposredno iskoristiti statistika $\overline{X}$, jer njezina razdioba vjerojatnosti $\mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ ovisi o nepoznatom parametru σ^2. Stoga

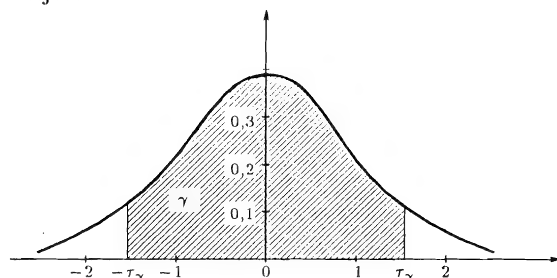
se prirodno nameće ideja da se pronađe neka statistika koja neće ovisiti o σ^2. U tu svrhu može se ovako zaključiti: Na temelju svojstva uzoračke aritmetičke sredine $\overline{X}$, izraženog relacijom (52) u VI.4, i formula (62) i (63) iz IV.6, proizlazi da $\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1)$. Tu se, međutim, još uvijek nalazi nepoznato σ, pa se nameće ideja da se σ^2 zamijeni nepristranim procjeniteljem S^2 (v. VI.4), čime se dobiva statistika

$$T = \frac{\overline{X}-\mu}{S}\sqrt{n} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}\sqrt{\frac{n-1}{\frac{n-1}{\sigma^2}S^2}}. \tag{20}$$

Iz (20) se vidi da se T može zapisati u obliku $T = Z\sqrt{\dfrac{n-1}{U}}$, gdje $Z = \dfrac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$ i $U = \dfrac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$, što je uočeno već u VI.4, formula (53). Na temelju, pak, onoga što je rečeno u točki 7. iz V.6, zaključuje se:

> Statistika T ima Studentovu ili t-razdiobu sa $n-1$ stupnjeva slobode. Piše se
> $$T \sim \mathrm{t}(n-1).$$

Time smo dobili statistiku koja ovisi samo o parametru μ, a ne i o nepoznatom parametru σ^2, dok njezina razdioba vjerojatnosti ovisi samo o veličini uzorka n ($n \geq 2$). U točki 7. iz V.6. također je navedeno, a vidi se i iz sl. 9, da je Studentova razdioba simetrična s obzirom na ishodište, pa se može zahtijevati da se odredi takvo $\tau_\gamma \geq 0$ da vrijedi



Slika 9. Skica krivulje Studentove razdiobe

$$\mathrm{P}(-\tau_\gamma < T < \tau_\gamma) = \gamma, \tag{21}$$

odnosno

$$\mathrm{P}\left(-\tau_\gamma < \frac{\overline{X}-\mu}{S}\sqrt{n} < \tau_\gamma\right) = \mathrm{P}\left(\overline{X} - \tau_\gamma\frac{S}{\sqrt{n}} < \mu < \overline{X} + \tau_\gamma\frac{S}{\sqrt{n}}\right) = \gamma. \tag{22}$$

Uspoređivanjem (4) i (22) zaključuje se da su statistike

$$G_{11} = \overline{X} - \tau_\gamma\frac{S}{\sqrt{n}}, \quad G_{12} = \overline{X} + \tau_\gamma\frac{S}{\sqrt{n}} \tag{23}$$

slučajni rubovi intervala povjerenja pouzdanosti γ za parametar μ normalne razdiobe $\mathrm{N}(\mu,\sigma^2)$ u uvjetima nepoznate varijance σ^2.

> Prema tome, praktično određivanje odgovarajućeg intervala povjerenja $\langle g_{11}, g_{12}\rangle$ sastoji se u tome da se, za zadano n i γ, najprije primjenom tablice za Studentovu razdiobu (v. tabl. V. u Dodatku), odredi pripadno τ_γ, a zatim se na danom nizu podataka $x_1,\ldots,x_n$ izračunaju vrijednosti
> $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 \tag{24}$$
> statistika $\overline{X}$ i S^2. Rubovi intervala povjerenja pouzdanosti γ za parametar μ glase
> $$g_{11} = \overline{x} - \tau_\gamma\frac{s}{\sqrt{n}}, \quad g_{12} = \overline{x} + \tau_\gamma\frac{s}{\sqrt{n}}. \tag{25}$$

Označi li se sa G_n f.r.v. Studentove razdiobe $\mathrm{t}(n)$ sa n stupnjeva slobode, onda je G_n^{-1} njoj inverzna funkcija, pa iz (21) proizlazi da je $\tau_\gamma = \mathrm{G}_{n-1}^{-1}\left(\dfrac{1+\gamma}{2}\right)$.

Tablica 4.

n	τ_γ		
	$\gamma = 0{,}90$	$\gamma = 0{,}95$	$\gamma = 0{,}99$
2	6,31	12,71	63,66
3	2,92	4,30	9,92
4	2,35	3,18	5,84
5	2,13	2,78	4,60
10	1,83	2,26	3,25
15	1,76	2,14	2,98
20	1,73	2,09	2,86
25	1,71	2,06	2,80
30	1,70	2,04	2,76
40	1,68	2,02	2,70
60	1,67	2,00	2,66
120	1,66	1,98	2,62
∞	1,65	1,96	2,58

Ako se, na primjer, na uzorku veličine $n = 25$ dobije aritmetička sredina $\overline{x} = 12$ i korigirana uzoračka varijanca $s^2 = 4$ i ako se zahtijeva pouzdanost $\gamma = 0{,}95$ pri procjeni nepoznatog očekivanja μ, onda će se najprije iz priložene tabl. 4. odčitati odgovarajuće $\tau_\gamma = 2{,}06$, a zatim, prema (25), izračunati rubovi intervala povjerenja

$$g_{11} = 12 - 2{,}06\cdot\frac{2}{\sqrt{25}} = 11{,}18, \quad g_{12} = 12 + 2{,}06\cdot\frac{2}{\sqrt{25}} = 12{,}82.$$

Usporede li se te vrijednosti s odgovarajućim vrijednostima rubova intervala povjerenja iz 1. primjera vidi se da je sada dobiven nešto širi interval povjerenja pouzdanosti 95 % za nepoznato očekivanje μ normalne razdiobe. To je posve razumljivo, jer su u 1. primjeru rubovi određeni uz poznatu varijancu $\sigma^2 = 4$, dok je ovdje primijenjena procjena $s^2 = 4$ za nepoznatu varijancu.

Iz (23) se razabire da je

$$\Delta_1 = |G_{12} - G_{11}| = 2\tau_\gamma \frac{S}{\sqrt{n}} \tag{26}$$

širina intervala povjerenja pouzdanosti γ za parametar μ normalne razdiobe. Iz (26) se vidi da Δ_1 ovisi o statistici S, tj. Δ_1 je slučajna varijabla, za razliku od 1. primjera, gdje širina intervala povjerenja nije bila slučajna varijabla.

Usporedbom tabl. 1. i 4. vidi se da je τ_γ uvijek veće od odgovarajućeg z_γ, a to znači da uz istu pouzdanost γ imamo širi interval povjerenja (veću grešku aproksimacije) kada se za varijancu σ^2 uzima procjena s^2 nego kada je varijanca unaprijed poznata. Tablica 4. pokazuje da je, za malene n, ta razlika vrlo značajna, dok se s porastom veličine uzorka n ona sve više smanjuje i za $n = \infty$ više nema razlike u granicama intervala povjerenja za nepoznato očekivanje μ normalne razdiobe izračunane na temelju (11) ili na temelju (23).

Iz (26) se razabire i to da vrijednost $\delta_1 = 2\tau_\gamma \frac{s}{\sqrt{n}}$ širine intervala povjerenja za parametar μ, osim što ovisi o vrijednosti s statistike S, ovisi i o neslučajnom faktoru $d_1 = \frac{2\tau_\gamma}{\sqrt{n}}$, pa se može reći da je δ_1 proporcionalno vrijednosti korigirane uzoračke standardne devijacije s koeficijentom proporcionalnosti d_1 koji ovisi o pouzdanosti γ i veličini uzorka n. Ta se ovisnost može konkretnije spoznati pogleda li se tabl. 5.

Tablica 5.

n	d_1		
	$\gamma = 0{,}90$	$\gamma = 0{,}95$	$\gamma = 0{,}99$
2	8,92	17,97	90,03
5	1,90	2,49	4,11
10	1,16	1,43	2,06
20	0,77	0,93	1,28
40	0,53	0,64	0,85
60	0,43	0,52	0,69
120	0,30	0,36	0,48

Ostalo je još da se razmotri i zadatak pojedinačnog određivanja intervala povjerenja za nepoznatu varijancu σ^2 normalne razdiobe. U tu svrhu iskoristit će se statistika $U = \frac{n-1}{\sigma^2} S^2$ za koju je u VI.4. (formula (53)) rečeno da joj pripada hikvadrat-razdioba sa $n-1$ stupnjeva slobode. Statistika U, dakle, ovisi samo o nepoznatom parametru σ^2, a njezina vjerojatnosna razdioba samo o veličini uzorka n ($n \geq 2$), pa je stoga prikladna za navedenu svrhu. Hikvadrat-razdioba, kako je opisano IV.5, nije simetrična pa će se zahtijevati da se odrede pozitivni brojevi u_1 i u_2 (v. sl. 10), tako da vrijedi

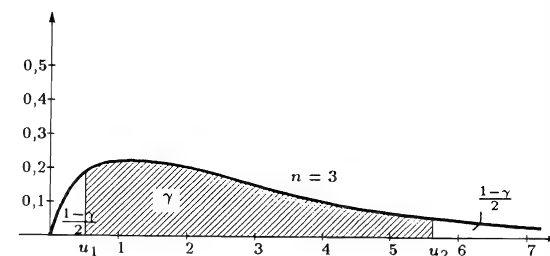
$$\mathrm{P}(U \leq u_1) = \frac{1-\gamma}{2}, \quad \mathrm{P}(U \geq u_2) = \frac{1-\gamma}{2}, \tag{27}$$

iz čega proizlazi da također vrijedi

$$\mathrm{P}(u_1 < U < u_2) = \gamma,$$

odnosno

$$\mathrm{P}\left(u_1 < \frac{n-1}{\sigma^2} S^2 < u_2\right) = \mathrm{P}\left(\frac{n-1}{u_2} S^2 < \sigma^2 < \frac{n-1}{u_1} S^2\right) = \gamma. \tag{28}$$



Slika 10. Krivulja hikvadrat-razdiobe i skica geometrijskog značenja pouzdanosti γ

Uspoređivanjem (4) i (28) vidi se da su statistike

$$G_{21} = \frac{n-1}{u_2} S^2, \quad G_{22} = \frac{n-1}{u_1} S^2 \tag{29}$$

slučajni rubovi intervala povjerenja pouzdanosti γ za parametar σ^2 normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$.

Označi li se s H_n f.r.v. hikvadrat-razdiobe $\chi^2(n)$ sa n stupnjeva slobode, onda se (27) može pisati kao

$$H_{n-1}(u_1) = \frac{1-\gamma}{2}, \quad 1 - H_{n-1}(u_2) = \frac{1-\gamma}{2},$$

iz čega slijedi da je

$$u_1 = H_{n-1}^{-1}\left(\frac{1-\gamma}{2}\right), \quad u_2 = H_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right).$$

Konkretne vrijednosti u_1 i u_2 mogu se izračunati pomoću odgovarajuće tablice za hikvadrat-razdiobu (v. tabl. VI. u Dodatku). U tabl. 6. navedene su vrijednosti za u_1 i u_2 u ovisnosti o nekim veličinama uzorka n i uobičajenih pouzdanosti γ.

Tablica 6.

n	$\gamma = 0{,}90$		$\gamma = 0{,}95$		$\gamma = 0{,}99$	
	u_1	u_2	u_1	u_2	u_1	u_2
2	0,004	3,84	0,001	5,02	0,0004	7,88
3	0,103	5,99	0,051	7,38	0,010	10,6
4	0,352	7,81	0,216	9,35	0,072	12,8
5	0,711	9,49	0,484	11,1	0,207	14,9
10	3,33	16,9	2,70	19,0	1,73	23,6
15	6,57	23,7	5,63	26,1	4,07	31,3
20	10,1	30,1	8,91	32,9	6,84	38,6
25	13,8	36,4	12,4	39,4	9,89	45,6
30	17,7	42,6	16,0	45,7	13,1	52,3
50	33,6	66,4	30,4	69,6	24,2	75,8
100	76,8	123	72,3	128	63,5	136

Praktično određivanje intervala povjerenja zadane pouzdanosti γ za varijancu σ^2 normalne razdiobe na temelju niza mjerenja $x_1, \ldots, x_n$ sastoji se u tome da se najprije, primjenom odgovarajuće tablice, odrede vrijednosti u_1 i u_2, zatim se izračuna vrijednost $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ statistike S^2 na danom nizu mjerenja, čime je omogućeno da se konačno izračunaju i rubovi intervala povjerenja

$$g_{21} = \frac{n-1}{u_2}s^2, \quad g_{22} = \frac{n-1}{u_1}s^2. \tag{30}$$

Ako je, na primjer, na uzorku veličine $n = 25$ dobivena vrijednost uzoračke korigirane varijance $s^2 = 4$, onda će interval povjerenja pouzdanosti $\gamma = 0{,}95$ za nepoznatu varijancu σ^2 imati rubove

$$g_{21} = \frac{24}{39{,}4} \cdot 4 = 2{,}44, \quad g_{22} = \frac{24}{12{,}4} \cdot 4 = 7{,}74.$$

To znači da izmjereni podaci sugeriraju da s pouzdanošću od 95 % možemo jamčiti da interval $\langle 2{,}44; 7{,}74\rangle$ pokriva nepoznatu varijancu promatrane slučajne varijable, uz pretpostavku da joj pripada normalna razdioba.

Iz (29) se razabire da je širina intervala povjerenja za varijancu σ^2 slučajna varijabla

$$\Delta_2 = |G_{22} - G_{21}| = (n-1)S^2\left(\frac{1}{u_1} - \frac{1}{u_2}\right), \tag{31}$$

pa se vidi da vrijednost δ_2 širine intervala povjerenja za parametar σ^2, osim što ovisi o vrijednosti s^2 korigirane uzoračke varijance, ovisi i o neslučajnom faktoru

$$d_2 = (n-1)\left(\frac{1}{u_1} - \frac{1}{u_2}\right).$$

Pogledom na tabl. 7. moguće je konkretnije uočiti ovisnost između d_2, n i γ.

Tablica 7.

n	d_2		
	$\gamma = 0{,}90$	$\gamma = 0{,}95$	$\gamma = 0{,}99$
2	250	1 000	2 500
5	5,2	7,9	19
10	2,2	2,9	4,8
20	1,2	1,6	2,3
30	0,96	1,2	1,7
50	0,72	0,91	1,4
100	0,49	0,59	0,83

Iz tabl. 7. vidi se da se s malim uzorcima ne mogu postići male greške pri intervalnoj procjeni nepoznate varijance σ^2 normalne razdiobe $N(\mu, \sigma^2)$. Tek s uzorcima koji se sastoje od stotinjak i više mjerenja dobivaju se intervali povjerenja čija je širina manja od vrijednosti korigirane uzoračke varijance s^2.

Zanimljivo je primjetiti da širina δ_1 intervala povjerenja za parametar μ ne ovisi o procjeni $\overline{x}$ parametra μ, već samo o procjeni s^2 (zapravo s) parametra σ^2. Također i širina δ_2 intervala povjerenja za parametar σ^2 ovisi o procjeni s^2. To je donekle i razumljivo jer varijanca σ^2 karakterizira raspršenje, tj. slučajnost u teorijskoj razdiobi vjerojatnosti kojom je izražena statistička zakonitost pri mjerenju slučajne varijable X, a s^2 karakterizira raspršenje u izmjerenom nizu podataka $x_1, \ldots, x_n$. Stoga je σ^2 glavni izvor nepouzdanosti pri procjeni parametara μ i σ^2, pa je razumljivo da će σ^2, odnosno odgovarajuća procjena s^2, bitno utjecati na veličinu greške procjene, tj. na širinu pripadnog intervala povjerenja.

Uspoređivanjem tabl. 5. i 7. također se vidi da je mnogo teže postići zadanu tolerantnu grešku procjene, izraženu faktorom proporcionalnosti d_1, odnosno d_2, pri procjeni varijance σ^2 nego pri procjeni očekivanja μ normalne razdiobe $N(\mu, \sigma^2)$.

Tako se, na primjer, iz tabl. 5. vidi da se s uzorcima od četrdesetak mjerenja ($n \approx 40$) već postiže $d_1 < 1$, tj. greška pri procjeni nepoznatog parametra μ nekom vrijednošću iz pripadnog intervala povjerenja ne premašuje uzoračku korigiranu standardnu devijaciju s. Da bi se postigla analogna točnost ($d_2 < 1$) pri procjeni varijance σ^2, iz tabl. 7. se vidi da treba uzeti stotinjak mjerenja ($n \approx 100$).

Polazeći od (19), (23) i (29) sada se može odrediti i simultani interval povjerenja za nepoznati vektorski parametar $\boldsymbol{t} = (\mu, \sigma^2)$ normalne razdiobe $N(\mu, \sigma^2)$. Vrijedi, naime

$$P\left(\overline{X} - \tau_\gamma\frac{S}{\sqrt{n}} < \mu < \overline{X} + \tau_\gamma\frac{S}{\sqrt{n}}, \ \frac{n-1}{u_2}S^2 < \sigma^2 < \frac{n-1}{u_1}S^2\right) \geq 1 - 2(1-\gamma),$$

što, prema (15), znači da je pravokutnik

$$\boldsymbol{I}_2 = \left\langle \overline{x} - \tau_\gamma\frac{s}{\sqrt{n}}, \overline{x} + \tau_\gamma\frac{s}{\sqrt{n}}\right\rangle \times \left\langle \frac{n-1}{u_2}s^2, \frac{n-1}{u_1}s^2\right\rangle \subseteq \mathbf{R}^2 \tag{32}$$

simultani interval povjerenja pouzdanosti $\gamma' = 1 - 2(1-\gamma)$ za vektorski parametar $\boldsymbol{t} = (\mu, \sigma^2)$.

Posebno, ako je na uzorku veličine $n = 25$ dobiveno $\overline{x} = 12$ i $s^2 = 4$, onda je $\boldsymbol{I}_2 = \langle 11{,}18 \ ; \ 12{,}82\rangle \times \langle 2{,}44 \ ; \ 7{,}74\rangle$ simultani interval povjerenja pouzdanosti

$\gamma' = 1-2(1-0{,}95) = 0{,}90$, što proizlazi iz maloprije izračunanih intervala povjerenja pouzdanosti $\gamma = 0{,}95$ za parametar μ i σ^2 pojedinačno.

3. Intervali povjerenja pri velikim uzorcima

Dosadašnja razmatranja pokazala su da je za određivanje intervala povjerenja zadane pouzdanosti γ za nepoznati parametar t, u smislu definicijske formule (4), potrebno definirati određenu statistiku $\widehat{T} = \widehat{T}_n = h(X_1,\ldots,X_n)$, koja je na određeni način povezana s nepoznatim parametrom t i kojoj pripada odgovarajuća razdioba vjerojatnosti. U svim razmotrenim primjerima bilo je relativno lako odrediti tu razdiobu vjerojatnosti, što je ključni moment pri rješavanju problema određivanja intervala povjerenja. Međutim, to uvijek nije tako, jer često postoje situacije kada tu razdiobu vjerojatnosti nije moguće jednostavno izraziti, što parakticki onemogućuje da se na taj način dođe do traženog intervala povjerenja. Stoga je nužno uvesti dodatne pretpostavke da bi se došlo do jednostavnijeg rješenja. Jedna od takvih pretpostavki jest da se raspolaže velikim uzorkom, što omogućuje primjenu asimptotskih svojstava statistike $\widehat{T}_n$.

Ako je $\widehat{T}_n$ asimptotski normalan i nepristran procjenitelj nepoznatog parametra t, onda se, prema relaciji (89) iz VI.8, može uzeti da približno vrijedi

$$\widehat{T}_n \sim \mathrm{N}(t, R_n(t)), \tag{33}$$

gdje je $R_n(t) = \mathrm{V}[\widehat{T}_n]$. To omogućuje da se napiše

$$\mathrm{P}(t - z_\gamma\sqrt{R_n(t)} < \widehat{T}_n < t + z_\gamma\sqrt{R_n(t)}) = \gamma, \tag{34}$$

gdje je z_γ definirano u (10) (v. tabl. 1).

Ako je moguće riješiti po t nejednadžbe

$$\begin{aligned} t - z_\gamma\sqrt{R_n(t)} &< \widehat{T}_n \\ t + z_\gamma\sqrt{R_n(t)} &> \widehat{T}_n, \end{aligned} \tag{35}$$

tako da se dobiju nejednadžbe

$$\begin{aligned} t &> G_1 \\ t &< G_2, \end{aligned} \tag{36}$$

onda su statistike G_1 i G_2 slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznati parametar t, jer je iz (34), (35) i (36) očigledno da vrijedi

$$\mathrm{P}(G_1 < t < G_2) = \gamma.$$

Problem se, dakle, praktički sveo na rješavanje sustava nejednadžbi (35), što dakako uvijek neće biti moguće. U tom slučaju i uz pretpostavku da je $\widehat{T}_n$ i konzistentan procjenitelj (v. VI.8) za parametar t, što znači da se za veliko n može uzeti $R_n(t) = R_n(\widehat{t}_n)$, jednadžba (34) postaje

$$\mathrm{P}\Big(t - z_\gamma\sqrt{R_n(\widehat{T}_n)} < \widehat{T}_n < t + z_\gamma\sqrt{R_n(\widehat{T}_n)}\Big) = \gamma,$$

odnosno

$$\mathrm{P}\Big(\widehat{T}_n - z_\gamma\sqrt{R_n(\widehat{T}_n)} < t < \widehat{T}_n + z_\gamma\sqrt{R_n(\widehat{T}_n)}\Big) = \gamma. \tag{37}$$

Uspoređivanjem (4) i (37) odmah se vidi da su

$$G_1 = \widehat{T}_n - z_\gamma\sqrt{R_n(\widehat{T}_n)}, \quad G_2 = \widehat{T}_n + z_\gamma\sqrt{R_n(\widehat{T}_n)}. \tag{38}$$

slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznati parametar t.

Ako je $\widehat{T}_n$ još i *asimptotski najefikasniji procjenitelj* (v. VI.7) za parametar t, tj. vrijedi

$$\lim_{n\to\infty} e(t) = \lim_{n\to\infty} \frac{1}{n\mathrm{V}[\widehat{T}_n]I(t)} = 1,$$

gdje je $I(t)$ Fisherova informacija, onda se za veliko n može približno uzeti da je $R_n(t) = \mathrm{V}[\widehat{T}_n] \approx \dfrac{1}{nI(t)}$, pa (38) postaje

$$G_1 = \widehat{T}_n - \frac{z_\gamma}{\sqrt{nI(\widehat{T}_n)}}, \quad G_2 = \widehat{T}_n + \frac{z_\gamma}{\sqrt{nI(\widehat{T}_n)}}. \tag{39}$$

Širina intervala povjerenja slučajna je varijabla

$$\Delta = |G_2 - G_1| = 2z_\gamma\sqrt{R_n(\widehat{T}_n)}\,.$$

3. primjer

Problem određivanja intervala povjerenja zadane pouzdanosti γ za nepoznati parametar α eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$ može se egzaktno riješiti (v. zad. 4) primjenom statistike $W = n\alpha\overline{X}$, gdje je n veličina uzorka, a $\overline{X}$ uzoračka aritmetička sredina. Na temelju poznatog rezultata iz točke 4. u V.6. zaključuje se da $W \sim \mathrm{G}(1,n)$. Ako se odrede vrijednosti $x_1, x_2 \in \mathbf{R}$ ($x_1 < x_2$) tako da vrijedi $\mathrm{P}(W \le x_1) = \mathrm{P}(W \ge x_2) = \dfrac{1-\gamma}{2}$, što povlači da je $\mathrm{P}(x_1 < W < x_2) = \gamma$, odnosno

$$\mathrm{P}\Big(\frac{x_1}{n\overline{X}} < \alpha < \frac{x_2}{n\overline{X}}\Big) = \gamma,$$

onda se odmah razabire da su

$$G_1 = \frac{x_1}{n\overline{X}}, \quad G_2 = \frac{x_2}{n\overline{X}} \tag{40}$$

slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznati parametar α eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$.

Nezgoda je, međutim, u tome što za određivanje x_1 i x_2 treba poznavati f.r.v. F_n gama-razdiobe $G(1,n)$, odnosno njoj inverznu funkciju F_n^{-1}, jer je $x_1 = F_n^{-1}\left(\frac{1-\gamma}{2}\right)$ i $x_2 = F_n^{-1}\left(\frac{1+\gamma}{2}\right)$, pa bi trebalo imati odgovarajuće tablice, što nije uobičajeno.

Ako je n dovoljno veliko, onda se mogu primijeniti formule (35), gdje će se uzeti da je $\widehat{T}_n = \overline{X}$, jer je poznato (v. 6. primjer u VI.3) da je $\overline{X}$ nepristran, konzistentan i asimptotski normalan procjenitelj za parametar $t = \alpha_0 = \frac{1}{\alpha}$ i $R_n(t) = \frac{1}{n}t^2$, pa (35) postaje

$$\frac{1}{\alpha}\left(1 - z_\gamma\frac{1}{\sqrt{n}}\right) < \overline{X}$$
$$\frac{1}{\alpha}\left(1 + z_\gamma\frac{1}{\sqrt{n}}\right) > \overline{X},$$

odnosno

$$\alpha > \frac{1}{\overline{X}}\left(1 - z_\gamma\frac{1}{\sqrt{n}}\right)$$
$$\alpha < \frac{1}{\overline{X}}\left(1 + z_\gamma\frac{1}{\sqrt{n}}\right),$$

iz čega proizlazi da su slučajni rubovi intervala povjerenja pouzdanosti γ za parametar α eksponencijalne razdiobe približno izraženi formulama

(41) $$G_1 = \left(1 - z_\gamma\frac{1}{\sqrt{n}}\right)\frac{1}{\overline{X}}, \quad G_2 = \left(1 + z_\gamma\frac{1}{\sqrt{n}}\right)\frac{1}{\overline{X}}.$$

Iz činjenice da je $\alpha > 0$ proizlazi da imaju smisla samo pozitivne vrijednosti za G_1, a to povlači da treba biti $n > z_\gamma^2$. Specijalno za $\gamma = 0{,}90$ treba biti $n > 2$, za $\gamma = 0{,}95$ treba $n > 3$, dok za $\gamma = 0{,}99$ treba $n > 6$. No, znamo da je opća pretpostavka za valjanost formula (41) da je n dovoljno veliko, tako da to, praktički gledano, i nisu značajna ograničenja.

Širina intervala povjerenja približno je izražena formulom

(42) $$\Delta = |G_2 - G_1| = 2z_\gamma\frac{1}{\sqrt{n}}\frac{1}{\overline{X}},$$

iz čega se vidi da ona opada s porastom veličine uzorka n kao $n^{-\frac{1}{2}}$, isto kao i u formulama (12) i (16).

4. primjer

Činjenica da je uzoračka aritmetička sredina $\overline{X}$ nepristran, konzistentan i asimptotski normalan procjenitelj za nepoznato očekivanje μ u modelu s dopuštenom klasom onih vjerojatnosnih razdioba koje imaju konačnu poznatu varijancu σ^2 (v. VI.2), može se iskoristiti za približno određivanje intervala povjerenja zadane pouzdanosti γ za nepoznato očekivanje μ.

Oslanjajući se na relaciju (26) iz VI.2 i stavljajući $t = \mu$ i $\widehat{T}_n = \overline{X}$, sustav nejednadžbi (35) postaje

$$\mu - z_\gamma\frac{\sigma}{\sqrt{n}} < \overline{X} \implies \mu < \overline{X} + z_\gamma\frac{\sigma}{\sqrt{n}}$$
$$\mu + z_\gamma\frac{\sigma}{\sqrt{n}} > \overline{X} \implies \mu > \overline{X} - z_\gamma\frac{\sigma}{\sqrt{n}},$$

pa se odmah vidi da su

(43) $$G_1 = \overline{X} - z_\gamma\frac{\sigma}{\sqrt{n}}, \quad G_2 = \overline{X} + z_\gamma\frac{\sigma}{\sqrt{n}},$$

slučajni rubovi intervala povjerenja pouzdanosti γ za parametar μ, dok je širina intervala povjerenja neslučajna veličina

(44) $$\delta = 2z_\gamma\frac{\sigma}{\sqrt{n}}.$$

Usporedbom (12) i (44) razabire se da je riječ o istim formulama, samo se mora imati na umu da (12) vrijedi za svaki $n \in \mathbf{N}$, ali za užu dopuštenu klasu razdioba, dok (44) vrijedi za velike n i mnogo širu klasu dopuštenih razdioba.

Kada je riječ o primjeni činjenice da je $\overline{X}$ asimptotski normalan procjenitelj za očekivanje μ, onda se pod dovoljno velikim n obično razumijeva $n \geq 30$, a nekada čak i $n \geq 15$.

Ako varijanca σ^2 nije poznata, onda se može poći od činjenice da niz slučajnih varijabli $Z_n = \frac{\overline{X} - \mu}{S}\sqrt{n}$ $(n \in \mathbf{N})$ konvergira po razdiobi (v. VI.8) slučajnoj varijabli $Z \sim N(0,1)$, što znači da za velike n približno vrijedi

$$P\left(-z_\gamma < \frac{\overline{X} - \mu}{S}\sqrt{n} < z_\gamma\right) = \gamma,$$

odnosno

$$P\left(\overline{X} - z_\gamma\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_\gamma\frac{S}{\sqrt{n}}\right) = \gamma,$$

iz čega proizlazi da su slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznato očekivanje μ

(45) $$G_1 = \overline{X} - z_\gamma\frac{S}{\sqrt{n}}, \quad G_2 = \overline{X} + z_\gamma\frac{S}{\sqrt{n}}.$$

Iz (43) i (45) vidi se da se rubovi intervala povjerenja za nepoznato očekivanje μ, uz nepoznatu varijancu σ^2, dobivaju iz rubova intervala povjerenja za μ uz poznato σ^2 tako da se σ u formulama (43) zamijeni vrijednošću korigirane uzoračke standardne

devijacije s. Usporede li se, pak, formule (23) i (45) vidi se da se one formalno razlikuju samo u tome što umjesto τ_γ stoji z_γ, a to upućuje na poznatu činjenicu da se za velike n ($n \geq 30$) Studentova razdioba $t(n)$ aproksimira standardnom normalnom razdiobom $N(0,1)$.

5. primjer

Svojstvo nepristranosti, konzistentnosti i asimptotske normalnosti uzoračke korigirane varijance S^2, kao procjenitelja za nepoznatu varijancu σ^2 u modelu s dopuštenom klasom onih vjerojatnosnih razdioba koje imaju konačni četvrti centralni moment μ_4, može se iskoristiti za približno određivanje intervala povjerenja zadane pouzdanosti γ za nepoznati parametar σ^2. Budući da postoji veza $\varepsilon = \frac{\mu_4}{\sigma^4} - 3$ (v. (15) u IV.2) između μ_4 i koeficijenta spljoštenosti ε, moguće je zadatak formulirati i tako da se traži interval povjerenja za σ^2 uz pretpostavku da je poznat koeficijent spljoštenosti ε.

Stavljajući $t = \sigma^2$ i $\widehat{T}_n = S^2$ može se pisati $\mu_4 = (\varepsilon + 3)t^2$, a formula (30) iz VI.2. tada ima oblik

$$R(t) = V[S^2] = \frac{1}{n}\left(\varepsilon + \frac{2n}{n-1}\right)t^2.$$

Sustav jednadžbi (35) u ovom slučaju postaje

$$(46)\qquad \begin{cases} \sigma^2\left(1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right) < S^2 \\ \sigma^2\left(1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right) > S^2 \end{cases}$$

Ako je $1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}} > 0$, onda se množenjem prve nejednadžbe sa $\left(1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1}$ dobiva

$$\sigma^2 < S^2\left(1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1},$$

dok se množenjem druge nejednadžbe sa $\left(1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1}$ dobiva

$$\sigma^2 > S^2\left(1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1},$$

što znači da su

$$(47)\qquad G_1 = S^2\left(1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1}, \quad G_2 = S^2\left(1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1}$$

slučajni rubovi intervala povjerenja pouzdanosti γ za parametar σ^2. Odgovarajuća širina intervala povjerenja slučajna je varijabla

$$(48)\qquad \Delta = \frac{2z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}}{1 - z_\gamma\left(\frac{\varepsilon}{n} + \frac{2}{n-1}\right)}S^2.$$

Ako je $1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}} = 0$, onda se iz (47) vidi da je $G_1 = \frac{1}{2}S^2$ i $G_2 = \infty$, pa se dobiva beskonačno širok interval povjerenja. Za $1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}} < 0$ sustav nejednadžbi (46) ima rješenje oblika $\sigma^2 > S^2\left(1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1}$, što znači da je $G_1 = S^2\left(1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}\right)^{-1}$ i $G_2 = \infty$, pa se i tada dobiva beskonačno širok interval povjerenja.

Promotri li se klasa vjerojatnosnih razdioba sa $\varepsilon = 0$, a takva je, na primjer, klasa normalnih razdioba, tada uz uvjet $1 - z_\gamma\sqrt{\frac{2}{n-1}} > 0$, odnosno $n > 2z_\gamma^2 + 1$, (47) postaje

$$(49)\qquad G_1 = S^2\left(1 + z_\gamma\sqrt{\frac{2}{n-1}}\right)^{-1}, \quad G_2 = S^2\left(1 - z_\gamma\sqrt{\frac{2}{n-1}}\right)^{-1},$$

a (48) postaje

$$(50)\qquad \Delta = \frac{2z_\gamma\sqrt{2(n-1)}}{n - 1 - 2z_\gamma^2}S^2.$$

Uvjet $n > 2z_\gamma^2 + 1$ praktički znači da za $\gamma = 0{,}90$ treba uzeti $n > 6$, za $\gamma = 0{,}95$ treba $n > 8$, dok za $\gamma = 0{,}99$ treba imati uzorak veličine $n > 14$ da bi se mogle primijeniti formule (49) i (50).

Primjedba

Problem određivanja intervala povjerenja za parametar σ^2 normalne razdiobe $N(\mu, \sigma^2)$ riješen je u VII.2. formulama (29) i (31). Uzme li se u obzir činjenica da se za velike n ($n > 30$) hikvadrat-razdioba $\chi^2(n)$ može aproksimirati normalnom razdiobom $N(n, 2n)$, tada se može približno uzeti da $U = \frac{n-1}{\sigma^2}S^2 \sim N(n-1, 2(n-1))$ i jednadžbe (27) mogu se zapisati u obliku

$$\Phi\left(\frac{u_1 - n + 1}{\sqrt{2(n-1)}}\right) = \frac{1-\gamma}{2}, \quad \Phi\left(\frac{u_2 - n + 1}{\sqrt{2(n-1)}}\right) = \frac{1+\gamma}{2},$$

odnosno

$$u_1 = \sqrt{2(n-1)}\,\Phi^{-1}\left(\frac{1-\gamma}{2}\right) + n - 1 = n - 1 - z_\gamma\sqrt{2(n-1)}$$

$$u_2 = \sqrt{2(n-1)}\,\Phi^{-1}\left(\frac{1+\gamma}{2}\right) + n - 1 = n - 1 + z_\gamma\sqrt{2(n-1)}$$

Uvrste li se te vrijednosti za u_1 i u_2 u (29), dobivaju se upravo formule (49), što pokazuje da stavljanjem $\varepsilon = 0$ u (47) dobivamo iste formule za slučajne rubove intervala povjerenja za nepoznatu varijancu σ^2 u općenitijem modelu (klasa dopuštenih razdioba sa zadanim koeficijentom spljoštenosti $\varepsilon = 0$), kao i u specijalnom modelu, gdje je klasa dopuštenih razdioba $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbf{R}, 0 < \sigma < \infty\}$. No, pritom valja stalno imati na umu da to vrijedi za velike uzorke ($n \geq 100$). Određeni uvid u pripadne širine intervala povjerenja za σ^2 u ovisnosti o veličini uzorka n i pouzdanosti γ može se dobiti pomoću tabl. 6. u VII.2.

4. Primjena Čebiševljeve nejednakosti

Značajno ograničenje u praktičnoj primjenjivosti opisanih metoda za određivanje intervala povjerenja sastoji se u pretpostavci da se polazi od statistike $\widehat{T}$, koja je u određenoj vezi s nepoznatim parametrom i, što je bitno, ima jednostavnu razdiobu vjerojatnosti, ili je asimptotski normalna, pa se izvođenje formula za rubove intervala povjerenja temelji na dobro proučenoj normalnoj razdiobi. Ako nijedna od tih pretpostavki nije ispunjena, onda se često problem može riješiti primjenom Čebiševljeve nejednakosti (v. IV.4).

Neka je, dakle, $\widehat{T}$ procjenitelj za parametar t u modelu s klasom $\mathcal{P} = \{P_t : t \in \Theta\}$ dopuštenih vjerojatnosnih razdioba i neka postoje konačni $E[\widehat{T}] = \mu_t$ i $V[\widehat{T}] = \sigma_t^2$, što omogućuje da se na slučajnu varijablu $\widehat{T}$ primijeni Čebiševljeva nejednakost, čime se dobiva

$$P(\mu_t - \lambda\sigma_t < \widehat{T} < \mu_t + \lambda\sigma_t) \geq 1 - \frac{1}{\lambda^2}, \tag{51}$$

gdje je $\lambda > 0$ proizvoljan realan broj.

Stavljajući $1 - \frac{1}{\lambda^2} = \gamma$ $(0 < \gamma < 1)$, odnosno

$$\lambda = \lambda_\gamma = \frac{1}{\sqrt{1-\gamma}}, \tag{52}$$

promotrimo sustav nejednadžbi

$$\mu_t - \lambda_\gamma\sigma_t < \widehat{T}, \qquad \mu_t + \lambda_\gamma\sigma_t > \widehat{T}. \tag{53}$$

Ako ga je moguće riješiti po t tako da se dobiju rješenja

$$t > G_1$$
$$t < G_2$$

onda (51) postaje ekvivalentno sa

$$P(G_1 < t < G_2) \geq \gamma, \tag{54}$$

a to, prema (3), znači da su G_1 i G_2 slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznati parametar t.

Odmah primijetimo da se nismo pozvali na relaciju (4), koja omogućuje dobivanje najužih intervala povjerenja, tako da će se ovim postupkom redovito dobivati intervali povjerenja koji su mnogo širi od najužih intervala povjerenja, koji su se dobivali već opisanim postupcima.

Ako bi se, na primjer, interval povjerenja za očekivanje μ, uz poznatu varijancu σ^2 (v. 4. primjer), određivao primjenom Čebiševljeve nejednakosti na procjenitelj $\widehat{T} = \overline{X}$, bilo bi $\mu_t = E[\overline{X}] = \mu$, $\sigma_t^2 = V[\overline{X}] = \frac{1}{n}\sigma^2$ i (53) postaje

$$\mu - \lambda_\gamma\frac{\sigma}{\sqrt{n}} < \overline{X} \implies \mu < \overline{X} + \lambda_\gamma\frac{\sigma}{\sqrt{n}}$$
$$\mu + \lambda_\gamma\frac{\sigma}{\sqrt{n}} > \overline{X} \implies \mu > \overline{X} - \lambda_\gamma\frac{\sigma}{\sqrt{n}},$$

tako da su

$$G_1 = \overline{X} - \lambda_\gamma\frac{\sigma}{\sqrt{n}}, \quad G_2 = \overline{X} + \lambda_\gamma\frac{\sigma}{\sqrt{n}}. \tag{55}$$

slučajni rubovi intervala povjerenja pouzdanosti γ za parametar μ.

Formule (55) vrlo su slične formulama (43). Razlika je samo u tome što umjesto z_γ stoji λ_γ. Vrijednosti z_γ, za određene vrijednosti pouzdanosti γ, navedene su u tabl. 1, a u tabl. 8. navest će se vrijednosti λ_γ, izračunane prema formuli (52).

Tablica 8.

γ	0,90	0,95	0,99
λ_γ	3,16	4,60	10

Širina intervala povjerenja, izračunana na temelju (55), jest

$$\delta = 2\lambda_\gamma\frac{\sigma}{\sqrt{n}}, \tag{56}$$

pa se usporedbom (44) i (56), te tabl. 1. i 8, vidi da se uz primjenu Čebiševljeve nejednakosti dobivaju nekoliko puta širi intervali povjerenja za nepoznato očekivanje μ, iste pouzdanosti γ, nego po metodi opisanoj u 4. primjeru.

Iako formule (55) i (56) vrijede za svaki $n \in \mathbf{N}$, za malene n praktički su beskorisne. Tako, na primjer, za $n = 4$ proizlazi da je širina intervala povjerenja $\delta = \lambda_\gamma\sigma$,

što za $\gamma = 0{,}95$ iznosi $\delta = 4{,}6\sigma$. To znači da se, bar sa 95 % pouzdanosti, a možda i više, može jamčiti da apsolutna greška pri aproksimaciji nepoznatog očekivanja μ uzoračkom aritmetičkom sredinom $\overline{x}$ neće premašiti, očigledno preveliku, vrijednost od gotovo pet standardnih devijacija.

Može se, prema tome, ustanoviti da interval povjerenja za nepoznato očekivanje μ, uz poznatu varijancu σ^2, ima smisla određivati primjenom Čebiševljeve nejednakosti samo kada veličina uzorka nije premalena, a ni prevelika $(5 < n < 20)$.

U primjedbi uz 5. primjer navedeno je da se formule za intervalnu procjenu nepoznate varijance σ^2, uz poznati koeficijent spljoštenosti ε, mogu primjenjivati za $n \geq 100$, jer tek tada dolazi do izražaja asimptotska normalnost procjenitelja S^2. Stoga se nameće ideja da se iskoristi Čebiševljeva nejednakost za određivanje intervala povjerenja za varijancu σ^2 kada je $n < 100$. Stavimo, dakle, $t = \sigma^2$ i $\widehat{T} = S^2$, pa je $\mu_t = \sigma^2$ i $\sigma_t^2 = \dfrac{1}{n}\left(\varepsilon + \dfrac{2n}{n-1}\right)\sigma^4$, što uvršteno u (53) daje

$$\text{(57)}\qquad \begin{aligned} \sigma^2\left(1-\lambda_\gamma\sqrt{\frac{\varepsilon}{n}+\frac{2}{n-1}}\right) &< S^2 \\ \sigma^2\left(1+\lambda_\gamma\sqrt{\frac{\varepsilon}{n}+\frac{2}{n-1}}\right) &> S^2. \end{aligned}$$

Nejednadžbe (57) formalno su identične nejednadžbama (46), samo što umjesto z_γ stoji λ_γ, pa će se i sve izvedene formule razlikovati samo u tome da se umjesto z_γ u formulama (47)-(50) stavi λ_γ, što će bitno utjecati na širinu odgovarajućih intervala povjerenja, tj. na točnost procjena.

Posebno je važno uočiti da se ovdje pojavljuje uvjet $n > 2\lambda_\gamma^2 + 1$, umjesto uvjeta $n > 2z_\gamma^2 + 1$ u 5. primjeru, što praktički znači da za $\gamma = 0{,}90$ treba biti $n > 26$, za $\gamma = 0{,}95$ treba $n > 43$, dok za $\gamma = 0{,}99$ treba imati uzorak veličine $n > 101$, da bi se mogle primijeniti izvedene formule.

5. Intervali povjerenja za vjerojatnost događaja

Mnoge praktične situacije zahtijevaju da se procijeni vjerojatnost p $(0 \leq p \leq 1)$ određenog događaja A, koji može i ne mora nastupiti u promatranome slučajnom eksperimentu ili slučajnoj pojavi. Na temelju n nezavisnih ponavljanja toga eksperimenta treba procijeniti vrijednost nepoznate vjerojatnosti p uočenog događaja A i odrediti pripadni interval povjerenja zadane pouzdanosti γ. Drugim riječima, treba procijeniti nepoznati parametar p i moguću grešku uz unaprijed zadani rizik da se donese pogrešan zaključak.

Tipičan primjer procjene vjerojatnosti događaja jest prognoziranje rezultata izbora za predsjednika države, ili za sastav parlamenta. Na temelju anketiranja određenog broja glasača, koji se odabiru po načelu slučajnosti iz skupa potencijalnih glasača, o tome za koga će glasovati, stručnjaci daju odgovarajuće prognoze o rezultatima budućih izbora, pri čemu navode i pouzdanost te prognoze.

Matematički model za opisani problem izgleda ovako: Pretpostavlja se da je broj svih mogućih glasača mnogo veći od broja n anketiranih glasača, tako da

se može smatrati da se uzorak veličine n uzima iz beskonačne populacije. Radi jednostavnosti uzimamo da anketno pitanje glasi: Da li ćete glasovati za kandidata K? Pretpostavlja se da u cijeloj populaciji glasača postoji određeni omjer p $(0 \leq p \leq 1)$ onih koji će glasovati za kandidata K. Broj p interpretira se kao vjerojatnost događaja A, da slučajno izabrani glasač glasuje za kandidata K. To se može egzaktno opisati pomoću slučajne varijable $X \sim \mathrm{B}(1,p)$. X je, dakle, slučajna varijabla Bernoullijeve razdiobe (v. IV.3), koja poprima vrijednost 0 (ne nastupa događaj A) s vjerojatnošću $1-p$ i vrijednost 1 (nastupa događaj A) s vjerojatnošću p. U teorijskom obliku zadatak glasi da se napravi intervalna procjena nepoznatog parametra p Bernoullijeve razdiobe na temelju n-članoga slučajnog uzorka $(X_1,\ldots,X_n)$, pri čemu je $X_i \sim \mathrm{B}(1,p)$, $i = 1,\ldots,n$.

Ranije je pokazano da je uzoračka aritmetička sredina $\overline{X} = \dfrac{1}{n}(X_1+\ldots+X_n)$, koja u ovom slučaju ima značenje relativne frekvencije događaja A u n-članome slučajnom uzorku, nepristran, konzistentan i asimptotski normalan procjenitelj za parametar p, pa će stoga $\overline{X}$ poslužiti kao osnova za određivanje intervala povjerenja zadane pouzdanosti γ za parametar p.

Za određivanje najužeg intervala povjerenja primjenom formula (5) i (6) trebalo bi upotrijebiti formulu (4) iz VI.1, kojom je definirana razdioba vjerojatnosti statistike $\overline{X}$. Tu se vidi da statistika $n\overline{X} \sim \mathrm{B}(n,p)$, pa se može postaviti zahtjev da se odrede c_1 i c_2 $(c_1 < c_2)$ tako da vrijedi

$$\text{(58)}\qquad \mathrm{P}(n\overline{X} \leq c_1) = \mathrm{P}(n\overline{X} \geq c_2) = \frac{1-\gamma}{2},$$

iz čega proizlazi da je

$$\text{(59)}\qquad \mathrm{P}(c_1 < n\overline{X} < c_2) = \gamma.$$

Iz činjenice da $n\overline{X} \sim \mathrm{B}(n,p)$ proizlazi da je

$$\text{(60)}\qquad \begin{aligned} \mathrm{P}\Big(\overline{X} \leq \frac{1}{n}c_1\Big) &= \sum_{j=0}^{c_1}\binom{n}{j}p^j(1-p)^{n-j} = \frac{1-\gamma}{2} \\ \mathrm{P}\Big(\overline{X} \geq \frac{1}{n}c_2\Big) &= \sum_{j=c_2}^{n}\binom{n}{j}p^j(1-p)^{n-j} = \frac{1-\gamma}{2}, \end{aligned}$$

što, bar načelno, omogućuje da se prva jednadžba riješi po c_1, a druga po c_2, čime se dobiva

$$\begin{aligned} c_1 &= c_1(p,n,\gamma) \\ c_2 &= c_2(p,n,\gamma), \end{aligned}$$

tj. c_1 i c_2 izraženi su u ovisnosti o p, n i γ. Sada još ostaje da se riješi po p sustav nejednadžbi

$$\text{(61)}\qquad \begin{aligned} c_1(p,n,\gamma) &< n\overline{X} \\ c_2(p,n,\gamma) &> n\overline{X}, \end{aligned}$$

čime se dobiva

$$G_1(n,\overline{X},\gamma) < p < G_2(n,\overline{X},\gamma),$$

i time su, načelno, određeni slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznatu vjerojatnost p, tj.

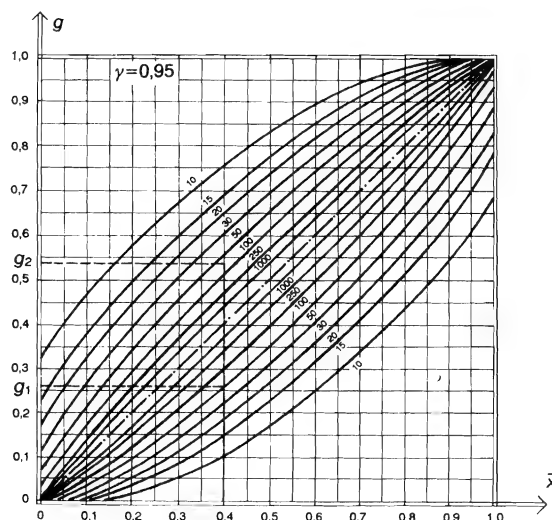
$$G_1 = G_1(n, \overline{X}, \gamma), \quad G_2 = G_2(n, \overline{X}, \gamma). \tag{62}$$

Numerički postupci za rješavanje jednadžbi (60) i sustava nejednadžbi (61) vrlo su složeni, tako da se zorni uvid u rješenje navedenog problema najbolje dobiva grafičkim prikazom rezultata, što je načinjeno na sl. 11. i 12.

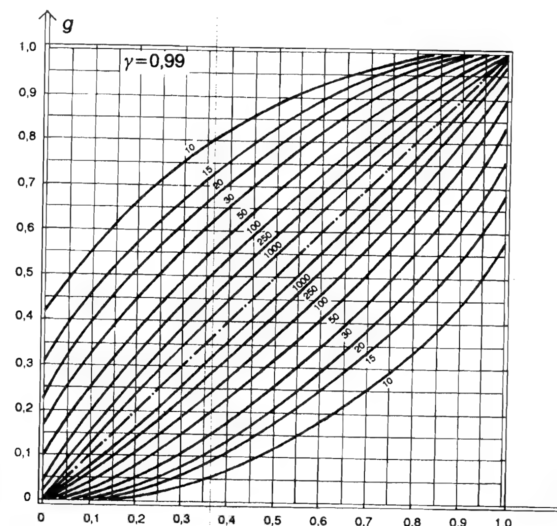
Ako se, na primjer, uzme $\gamma = 0{,}95$ i na uzorku veličine $n = 50$ dobije vrijednost relativne frekvencije promatranog događaja $\overline{x} = 0{,}4$, onda se na sl. 11. vidi da vertikalni pravac, kroz točku apscise 0,4, siječe ucrtane krivulje za $n = 50$ u točkama koje imaju ordinate $g_1 \approx 0{,}25$ i $g_2 \approx 0{,}55$, i to su upravo rubovi traženog intervala povjerenja. Konkretno, ako se anketiranjem 50 glasača želi prognozirati rezultat izbora i ako se ustanovi da je 20 glasača najavilo glasovati za kandidata K $\left(\overline{x} = \dfrac{20}{50} = 0{,}4\right)$, onda se, s pouzdanošću od 95 %, može jamčiti da bi kandidat K na općim izborima dobio između 25 % i 55 % glasova.

Ako se nekome čini da je to preširok interval povjerenja, onda će, dakako, morati povećati veličinu uzorka n. No, da bi se moglo egzaktnije zaključivati o vezi između veličine uzorka n i širine intervala povjerenja, trebalo bi naći formulu za širinu intervala povjerenja δ, kako je već ranije činjeno.

Budući da upravo provedeno razmatranje nije omogućilo dobivanje eksplicitnih formula za slučajne rubove G_1 i G_2 intervala povjerenja u ovisnosti o n, problem će se riješiti za velike n metodom opisanom u VII.3. Polazi se od činjenice da je uzoračka relativna frekvencija $\overline{X}$ asimptotski normalan procjenitelj za nepoznatu vjerojatnost p i da je $\mathrm{E}[\overline{X}] = p$, $\mathrm{V}[\overline{X}] = \dfrac{1}{n}p(1-p)$, pa ako se u (35) uvrsti $t = p$,



Slika 11. Intervali povjerenja za vjerojatnost p uz $\gamma = 0{,}95$



Slika 12. Intervali povjerenja za vjerojatnost p uz $\gamma = 0{,}99$

$\widehat{T}_n = \overline{X}$ i $R_n(t) = \dfrac{1}{n}p(1-p)$, dobiva se

$$\begin{aligned} p - z_\gamma\sqrt{\frac{1}{n}p(1-p)} &< \overline{X} \\ p + z_\gamma\sqrt{\frac{1}{n}p(1-p)} &> \overline{X}. \end{aligned} \tag{63}$$

Sustav nejednadžbi (63) ekvivalentan je nejednadžbi

$$(\overline{X} - p)^2 - \frac{1}{n}z_\gamma^2 p(1-p) < 0,$$

odnosno nejednadžbi

$$(n + z_\gamma^2)p^2 - (2n\overline{X} + z_\gamma^2)p + n\overline{X}^2 < 0. \tag{64}$$

Lijeva strana u (64) je kvadratni polinom u varijabli p, koji će poprimati negativne vrijednosti između svojih nul-točaka

$$(65)\qquad \begin{aligned} p_1 = G_1 &= \frac{\overline{X} + \frac{1}{2n}z_\gamma^2 - z_\gamma\sqrt{\frac{1}{n}\overline{X}(1-\overline{X}) + \frac{1}{4n^2}z_\gamma^2}}{1 + \frac{1}{n}z_\gamma^2}, \\ p_2 = G_2 &= \frac{\overline{X} + \frac{1}{2n}z_\gamma^2 + z_\gamma\sqrt{\frac{1}{n}\overline{X}(1-\overline{X}) + \frac{1}{4n^2}z_\gamma^2}}{1 + \frac{1}{n}z_\gamma^2}. \end{aligned}$$

Rješenje sustava nejednadžbi (63) po parametru p, koje je identično rješenju kvadratne nejednadžbe (64), prema tome glasi

$$G_1 < p < G_2.$$

To, pak, prema (36), znači da su G_1 i G_2, izraženi formulama (65), slučajni rubovi intervala povjerenja pouzdanosti γ za nepoznatu vjerojatnost p uočenog događaja A, pri čemu se, dakako, pretpostavlja da je n dovoljno veliko, tj. takvo da se binomna razdioba $\mathrm{B}(n,p)$ smije aproksimirati normalnom razdiobom $\mathrm{N}(np, np(1-p))$.

Ako bismo na već razmotreni konkretni primjer ($\gamma = 0{,}95$, $n = 50$) primijenili formule (65), dobili bismo konkretne rubove intervala povjerenja

$$g_1 = \frac{0{,}4 + 0{,}01 \cdot 1{,}96^2 - 1{,}96\sqrt{0{,}02 \cdot 0{,}4 \cdot 0{,}6 + 0{,}0001 \cdot 1{,}96^2}}{1 + 0{,}02 \cdot 1{,}96^2} \approx 0{,}276 = 27{,}6\,\%,$$

$$g_2 \approx 0{,}538 = 53{,}8\,\%,$$

čemu odgovara širina intervala povjerenja $\delta \approx 0{,}262$.

Usporede li se ti rezultati s onima dobivenim na temelju sl. 11, vidimo da razlike i nisu jako velike, zbog toga što je $n = 50$ već dovoljno veliko za aproksimaciju binomne razdiobe odgovarajućom normalnom razdiobom.

Za $n > 100$ mogu se u (65) zanemariti članovi $\frac{1}{2n}z_\gamma^2$ i $\frac{1}{4n^2}z_\gamma^2$, pa se dobivaju jednostavnije formule

$$(66)\qquad G_1 = \overline{X} - z_\gamma\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})},\quad G_2 = \overline{X} + z_\gamma\sqrt{\frac{1}{n}\overline{X}(1-\overline{X})},$$

na temelju kojih se dobiva i jednostavna formula

$$(67)\qquad \Delta = \frac{2}{\sqrt{n}}z_\gamma\sqrt{\overline{X}(1-\overline{X})},$$

za širinu Δ intervala povjerenja. Za svaku vrijednost $\overline{x}$ $(0 \le \overline{x} \le 1)$ statistike $\overline{X}$ vrijedi da je $\sqrt{\overline{x}(1-\overline{x})} \le \frac{1}{2}$ i stoga je vrijednost širine intervala povjerenja

$$(68)\qquad \delta \le \frac{1}{\sqrt{n}}z_\gamma = \frac{d_0}{2}.$$

Iz (68) se zaključuje da za postizanje unaprijed zadane širine δ (ili manje) intervala povjerenja pouzdanosti γ, pri procjeni nepoznate vjerojatnosti p, treba uzeti uzorak čija veličina n zadovoljava nejednakost

$$(69)\qquad n \ge \frac{1}{\delta^2}z_\gamma^2.$$

Tako, na primjer, želi li se prognozirati rezultat glasovanja uz pouzdanost $\gamma = 0{,}95$ tako da greška (širina intervala povjerenja) ne bude veća od 10 %, trebat će anketirati bar $n = \frac{1}{0{,}1^2} \cdot 1{,}96 = 385$ glasača. Za prognozu iste pouzdanosti, ali uz tolerantnu grešku od 5 %, treba anketirati bar $n = \frac{1}{0{,}05^2}\,1{,}96 = 1\,537$ glasača.

6. Bayesovski intervali povjerenja

U VI.9. opisana je Bayesova metoda procjene nepoznatog parametra t, koja se primjenjuje kada istraživač, osim izmjerenih podataka $x_1, \ldots, x_n$ o slučajnoj varijabli X, čija razdioba vjerojatnosti pripada klasi $\mathcal{P} = \{P_t : t \in \Theta\}$ dopuštenih vjerojatnosnih razdioba, raspolaže i određenom apriornom vjerojatnosnom razdiobom Π, koja opisuje ponašanje parametra t kao slučajne varijable T. Neka je $Y = h(X_1, \ldots, X_n)$ određena statistika, pa se može promatrati slučajni vektor (T, Y), čija je razdioba vjerojatnosti određena razdiobom Π i uvjetnim razdiobama vjerojatnosti statistike Y za svako $t \in \Theta$. No, tada je određena i uvjetna razdioba vjerojatnosti slučajne varijable T uz uvjet da slučajna varijabla Y poprimi vrijednost y. Označimo sa T_y slučajnu varijablu kojoj pripada ta vjerojatnosna razdioba, pa se može postaviti zadatak da se odrede brojevi g_1 i g_2 $(g_1 < g_2)$, tako da vrijedi

$$(70)\qquad \mathrm{P}(g_1 < T_y < g_2) = \gamma,$$

gdje je γ $(0 < \gamma < 1)$ unaprijed zadan broj. Kaže se da je $\langle g_1, g_2\rangle$ *Bayesovski interval povjerenja pouzdanosti* γ za nepoznati parametar t. Govori se još da je interval $\langle g_1, g_2\rangle$ određen na temelju *aposteriorne vjerojatnosne razdiobe* nepoznatog parametra t, nastale pod utjecajem opaženih (izmjerenih) podataka $x_1, \ldots, x_n$.

Primijetimo da je očekivanje slučajne varijable T_y, tj. uvjetno očekivanje slučajne varijable T uz uvjet da je statistika Y poprimila vrijednost $y = h(x_1, \ldots, x_n)$, točkasta procjena za nepoznati parametar t u smislu Bayesove metode (v. VI.9).

Da bismo konkretno ilustrirali određivanje Bayesovskog intervala povjerenja vratimo se primjeru iz VII.5, gdje je nepoznati parametar $t = p$ označivao omjer (postotak) glasača koji bi na izborima glasovali za kandidata K. Kao apriorna razdioba nepoznatog parametra p može se uzeti, recimo, uniformna razdioba $\mathrm{U}(0,1)$, što praktički znači da subjektivna informacija o ishodu glasovanja za kandidata K ne omogućuje favoriziranje nekog omjera (postotka) iz intervala $\langle 0, 1\rangle$. Nakon anketiranja n potencijalnih glasača dobiven je niz $x_1, \ldots, x_n$ (nula i jedinica) i na temelju svega treba naći pripadni Bayesovski interval povjerenja zadane pouzdanosti γ.

Kao što smo već i ranije činili, uzet ćemo statistiku $\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$, koja ovdje ima značenje relativne frekvencije promatranog događaja (slučajno izabrani glasač daje svoj glas kandidatu K) i za koju znamo da je dobar procjenitelj za parametar p. Također znamo da slučajna varijabla $Y = n\overline{X}$ ima binomnu razdiobu $\mathrm{B}(n, p)$ $(0 < p < 1)$.

Sličnim razmatranjem kao u 12. primjeru iz VI.9. zaključujemo da slučajnom vektoru (T, Y) pripada dvodimenzionalna vjerojatnosna razdioba opisana funkcijom

$$f(t, k) = f_1(t)\,\mathrm{P}(Y = k/T = t),$$

gdje je f_1 f.g.v. uniformne razdiobe $\mathrm{U}(0, 1)$ i

$$\mathrm{P}(Y = k/T = t) = \binom{n}{k} t^k (1-t)^{n-k}, \quad k = 0, 1, \ldots, n,$$

tako da se konačno dobiva

$$(71) \qquad f(t, k) = \begin{cases} \binom{n}{k} t^k (1-t)^{n-k} & , \ za \ \ 0 < t < 1, \quad k = 0, 1, \ldots, n \\ 0 & , \ \text{inače.} \end{cases}$$

To omogućuje da se dobije

$$(72) \qquad \mathrm{P}(Y = k) = \binom{n}{k} \int_0^1 t^k (1-t)^{n-k}\,\mathrm{d}t = \frac{1}{n+1}, \quad k = 0, 1, \ldots, n,$$

što proizlazi iz činjenice da je

$$(73) \qquad (n+1)\binom{n}{k} t^k (1-t)^{n-k} = \frac{(n+1)!}{k!(n-k)!} t^k (1-t)^{n-k} = \\ = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha - 1} (1-t)^{\beta - 1},$$

tj. riječ je o f.g.v. beta-razdiobe s parametrima $\alpha = k + 1$ i $\beta = n - k + 1$ (v. IV.5), čiji integral, dakako, iznosi 1.

Iz (71) i (72) lako se dobiva uvjetna razdioba vjerojatnosti slučajne varijable T uz uvjet da je statistika Y poprimila vrijednost k. Pripadna f.g.v. glasi

$$(74) \qquad p_k(t) = \frac{f(t, k)}{\mathrm{P}(Y = k)} = \begin{cases} (n+1)\binom{n}{k} t^k (1-t)^{n-k} & , \text{za } 0 < t < 1 \\ 0 & , \text{inače.} \end{cases}$$

Odmah se vidi da je formulom (74) definirana beta-razdioba s parametrima $\alpha = k + 1$ i $\beta = n - k + 1$ i da će stoga rubovi g_1 i g_2, pripadnoga Bayesovskog intervala povjerenja pouzdanosti γ za nepoznati parametar $t = p$, biti određeni formulama

$$(75) \qquad \int_0^{g_1} p_k(t)\,\mathrm{d}t = \int_{g_2}^1 p_k(t)\,\mathrm{d}t = \frac{1-\gamma}{2}.$$

Izračunavanje integrala u (75) i rješavanje odgovarajućih jednadžbi po g_1 i g_2 može biti vrlo mukotrpan posao. Stoga se, umjesto određivanja najužeg intervala povjerenja, može primijeniti Čebiševljeva nejednakost za dobivanje približnog rješenja. Poznato je (v. (58) u IV.5), naime, da postoje jednostavne formule za očekivanje i varijancu beta-razdiobe, na temelju kojih proizlazi

$$\mu_y = \mathrm{E}[T_y] = \frac{k+1}{n+2}, \qquad \sigma_y^2 = \mathrm{V}[T_y] = \frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}.$$

Napiše li se Čebiševljeva nejednakost za slučajnu varijablu T_y i usporedi s relacijom (70), te imajući na umu značenje oznake λ_γ, definirane u (52), dobivaju se približni rubovi Bayesovskog intervala povjeranja pouzdanosti γ izraženi formulama

$$(76) \qquad g_1 = \mu_y - \lambda_\gamma \sigma_y, \quad g_2 = \mu_y + \lambda_\gamma \sigma_y.$$

Tako, na primjer, ustanovi li se na uzorku od $n = 50$ glasača da bi njih $k = 20$ glasovalo za kandidata K, onda se s pouzdanošću od 95 % može jamčiti da bi kandidat K na općim izborima dobio između $g_1 = \frac{21}{52} - 4{,}6 \cdot \sqrt{\frac{21 \cdot 31}{52^2 \cdot 53}} \approx 0{,}40 - 0{,}31 = 0{,}09 = 9\,\%$ i $g_2 \approx 0{,}40 + 0{,}31 = 0{,}71 = 71\,\%$ glasova.

Dobiven je, naravno, vrlo širok interval povjerenja, što je bilo i za očekivati imajući na umu sve ono što je rečeno u VII.4. glede primjene Čebiševljeve nejednakosti.

Za velike n i $k \approx \frac{n}{2}$ dobivena beta-razdioba može se aproksimirati normalnom razdiobom $N(\mu_y, \sigma_y^2)$, što omogućuje da se rubovi odgovarajućeg Bayesovskog intervala povjerenja izraze formulama

$$(77) \qquad g_1 = \mu_y - z_\gamma \sigma_y, \quad g_2 = \mu_y + z_\gamma \sigma_y.$$

U već spomenutome konkretnom primjeru dobili bismo $g_1 \approx 0{,}40 - 1{,}96 \cdot 0{,}067 = 0{,}27 = 27\,\%$ i $g_2 \approx 0{,}53 = 53\,\%$, što je već mnogo uži interval povjerenja. On je čak uži i od intervala povjerenja dobivenog u VII.5. bez pretpostavke o apriornoj razdiobi vjerojatnosti nepoznatog parametra.

Zadaci

1. Načinjeno je n nezavisnih mjerenja slučajne varijable X, za koju se pretpostavlja da ima normalnu razdiobu poznate varijance $\sigma^2 = 4$. Treba naći interval povjerenja pouzdanosti γ za nepoznato očekivanje μ, ako je vrijednost uzoračke sredine $\overline{x} = 0$ i:

a) $n = 10$, $\gamma = 0{,}95$, b) $n = 50$, $\gamma = 0{,}95$,
c) $n = 100$, $\gamma = 0{,}95$, d) $n = 10$, $\gamma = 0{,}99$,
e) $n = 50$, $\gamma = 0{,}99$, f) $n = 100$, $\gamma = 0{,}99$.

2. Kolika treba biti veličina uzorka n da bi se odredio interval povjerenja pouzdanosti $\gamma = 0{,}99$, čija je širina:

a) 3σ, b) 2σ, c) σ, d) $0{,}5\sigma$, e) $0{,}1\sigma$, f) $0{,}01\sigma$.
Pretpostavlja se da je slučajni uzorak uzet iz normalne razdiobe poznate varijance σ^2.

3. Nađite interval povjerenja pouzdanosti $\gamma = 0{,}99$ za parametar μ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$, kada je σ^2 poznato i kada se raspolaže samo jednim mjerenjem x_1 $(n = 1)$.

4. Izvedite formule za slučajne rubove G_1 i G_2 intervala povjerenja pouzdanosti γ za parametar α eksponencijalne razdiobe $\mathrm{Ex}(\alpha)$.
Uputa: Primijenite činjenicu da $\alpha X_i \sim \mathrm{Ex}(1)$ i da $\alpha \sum_{i=1}^{n} X_i \sim G(1, n)$.

5. Nađite formulu za očekivanje $\mathrm{E}[\Delta_2]$ širine intervala povjerenja pri procjeni nepoznate varijance σ^2 normalne razdiobe.
Uputa: Iskoristite činjenicu da $\dfrac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$.

6. Nađite približnu formulu za očekivanje $\mathrm{E}[\Delta_1]$ širine intervala povjerenja pri procjeni nepoznatog očekivanja μ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$ u uvjetima nepoznate varijance i uporedite dobiveni rezultat sa širinom intervala povjerenja u uvjetima poznate varijance.

7. Neka su $(X_1, \ldots, X_m)$ i $(Y_1, \ldots, Y_n)$ nezavisni slučajni uzorci za nezavisne slučajne varijable $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ i $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Izvedite formule za slučajne rubove G_1 i G_2 intervala povjerenja za veličinu $d_0 = \mu_1 - \mu_2$, uz pretpostavku da:

a) σ_1^2 i σ_2^2 su nepoznate veličine,
b) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ je nepoznato.

Uputa: U a) iskoristite činjenicu da statistika

$$Z = \frac{\overline{X} - \overline{Y} - d_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim \mathrm{N}(0, 1),$$

a u b) činjenicu da statistika

$$T = \frac{\overline{X} - \overline{Y} - d_0}{\overline{S}\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim \mathrm{t}(m + n - 2),$$

gdje je

$$\overline{S}^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m + n - 2},$$

a S_x^2 i S_y^2 su odgovarajuće korigirane uzoračke varijance.

8. Nađite simultani interval povjerenja pouzdanosti $\gamma = 0{,}90$ za nepoznati vektorski parametar $\boldsymbol{t} = (\mu, \sigma^2)$ normalne razdiobe, ako su na uzorku veličine $n = 30$ dobivene vrijednosti $\overline{x} = 5$ i $s^2 = 5{,}76$.

9. Neka je $n > 1$ i $(X_1, \ldots, X_n)$ slučajni uzorak za slučajnu varijablu $X \sim \mathrm{N}(\mu, \sigma^2)$, gdje su μ i σ^2 nepoznati parametri. Neka slučajna varijabla

X_{n+1} označuje n plus prvo nezavisno mjerenje slučajne varijable X. Nađite tzv. *interval proricanja* zadane pouzdanosti γ za vrijednost x_{n+1} slučajne varijable X_{n+1}.
Uputa: Dokažite najprije da

$$T = \frac{\overline{X} - X_{n+1}}{S}\sqrt{\frac{n}{n+1}} \sim \mathrm{t}(n-1).$$

10. Uz pretpostavku da slučajna varijabla X (tlačna čvrstoća betona) iz 5. primjera u I.4. ima $\mathrm{N}(\mu, \sigma^2)$ nađite intervale povjerenja za μ i σ^2 pouzdanosti:

a) $\gamma = 0{,}90$, b) $\gamma = 0{,}95$, c) $\gamma = 0{,}99$.

11. Predvidite, uz pouzdanost $\gamma = 0{,}90$, vrijednost slučajne varijable X (godišnja količina oborina) iz zad. 10. u I. pogl. za iduću godinu, primjenjujući rezultat iz zad. 9.

12. Nađite intervale povjerenja za očekivanje μ i varijancu σ^2 pouzdanosti $\gamma = 0{,}95$, uzimajući kao vrijednost slučajnog uzorka podatke iz:
a) zad. 11. u I. pogl., b) zad. 12. u I. pogl., c) zad. 13. u I. pogl.

13. Na temelju niza podataka iz 1. primjera u I.1. nađite interval povjerenja pouzdanosti $\gamma = 0{,}90$ za vjerojatnost p da učenik završi razred:

a) s odličnim uspjehom iz matematike,
b) s negativnom ocjenom iz matematike,
c) s pozitivnom ocjenom iz matematike.

14. Na temelju niza podataka iz 3. primjera u I.2. nađite interval povjerenja pouzdanosti $\gamma = 0{,}99$ za:

a) očekivani dnevni broj kvarova,
b) varijancu dnevnog broja kvarova,
c) vjerojatnost da dnevni broj kvarova bude veći od 10.

15. Na temelju niza podataka iz 4. primjera u I.3. nađite interval povjerenja pouzdanosti $\gamma = 0{,}95$ za vjerojatnost (postotak) slova A u tekstovima hrvatskog jezika.

16. Odredite interval povjerenja pouzdanosti $\gamma = 0{,}95$ za očekivanje μ i varijancu σ^2 na temelju niza podataka iz:
a) zad. 2. u I. pogl., b) zad. 3. u I. pogl., c) zad. 5. u I. pogl.

17. Odredite Bayesovski interval povjerenja pouzdanosti γ za nepoznati parametar iz:

a) 13. primjera u VI.9,
b) zad. 29. u VI. pogl.,
c) zad. 30. u VI. pogl.

Pregled važnijih intervala povjerenja

Pretpostavljena klasa razdioba	Parametar	Donji rub	Gornji rub	Primjedba
B(1, p)	p	$\overline{x} - z_\gamma \sqrt{\frac{1}{n}\overline{x}(1-\overline{x})}$	$\overline{x} + z_\gamma \sqrt{\frac{1}{n}\overline{x}(1-\overline{x})}$	$n \geq 40$
Po(λ)	λ	$\overline{x} - z_\gamma \sqrt{\frac{\overline{x}}{n}}$	$\overline{x} + z_\gamma \sqrt{\frac{\overline{x}}{n}}$	$n \geq 30$
Ex(α)	α	$\left(1 - z_\gamma \frac{1}{\sqrt{n}}\right)\frac{1}{\overline{x}}$	$\left(1 + z_\gamma \frac{1}{\sqrt{n}}\right)\frac{1}{\overline{x}}$	$n \geq 30$
U(0, t)	t	$\max(x_1, \ldots, x_n)$	$\frac{\max(x_1, \ldots, x_n)}{\sqrt[n]{1-\gamma}}$	najuži interval
N(μ, σ^2) σ^2 poznato	μ	$\overline{x} - z_\gamma \frac{\sigma}{\sqrt{n}}$	$\overline{x} + z_\gamma \frac{\sigma}{\sqrt{n}}$	najuži interval
N(μ, σ^2)	μ	$\overline{x} - \tau_s \frac{s}{\sqrt{n}}$	$\overline{x} + \tau_\gamma \frac{s}{\sqrt{n}}$	najuži interval
	σ^2	$\frac{n-1}{u_2}s^2$	$\frac{n-1}{u_1}s^2$	$u_1 = G_{n-1}^{-1}\left(\frac{1-\gamma}{2}\right)$, $u_2 = G_{n-1}^{-1}\left(\frac{1+\gamma}{2}\right)$, najuži interval
postoji konačna poznata varijanca σ^2	μ	$\overline{x} - z_\gamma \frac{\sigma}{\sqrt{n}}$	$\overline{x} + z_\gamma \frac{\sigma}{\sqrt{n}}$	$n \geq 15$
postoji konačna varijanca	μ	$\overline{x} - z_\gamma \frac{s}{\sqrt{n}}$	$\overline{x} + z_\gamma \frac{s}{\sqrt{n}}$	$n \geq 30$
postoji konačni poznati koeficijent spljoštenosti ε	σ^2	$\frac{s^2}{1 + z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}}$	$\frac{s^2}{1 - z_\gamma\sqrt{\frac{\varepsilon}{n} + \frac{2}{n-1}}}$	$\frac{\varepsilon}{n} + \frac{2}{n-1} < z_\gamma^2$, $n \geq 100$

VIII. Testiranje parametarskih hipoteza

1. Uvod u problematiku

Mnoge praktične situacije u vezi sa slučajnim pojavama zahtijevaju da se donesu odluke tipa DA ili NE. Tako, na primjer, pri praćenju procesa proizvodnje nekog proizvoda treba, na temelju rezultata mjerenja $x_1, \ldots, x_n$ relevantne veličine X, donijeti odluku o tome da li proces proizvodnje osigurava ili ne osigurava zahtijevanu kvalitetu. Pretpostavlja se, dakako, da veličina X, koja karakterizira kvalitetu pojedinog proizvoda (količina određenog sastojka, na primjer), ima slučajni karakter. Obično se smatra da je zahtijevana kvaliteta postignuta ako X ima unaprijed zadana svojstva (da se, na primjer, kreće u danim granicama, ili da ima zadanu srednju vrijednost i standardnu devijaciju i sl.).

Ako se, na primjer, smatra da proces proizvodnje osigurava zahtijevanu kvalitetu kada ne daje više od 10 % neispravnih proizvoda, onda se izjava: "Proces proizvodnje osigurava zahtijevanu kvalitetu", može iskazati i kao statistička hipoteza: "Slučajna varijabla X ima svojstvo da je $P(X \leq x_0) \geq 0{,}90$", gdje je x_0 određena "kritična vrijednost" slučajne varijable X koja ne smije biti premašena na proizvodu zadovoljavajuće kvalitete. Izjava: "Proces proizvodnje osigurava zahtijevanu kvalitetu", u nekoj drugoj prilici, može biti izražena nekom drugom statističkom hipotezom, recimo $E[X] = \mu_0$ i $V[X] = \sigma_0^2$ ili, na primjer, statističkom hipotezom $X \sim N(\mu_0, \sigma_0^2)$, gdje su μ_0 i σ_0^2 unaprijed zadane veličine.

Teorijski gledano, riječ je o tome da se, na temelju n-članog niza mjerenja slučajne varijable X, odnosno na temelju vrijednosti $(x_1, \ldots, x_n)$ slučajnog uzorka $(X_1, \ldots, X_n)$, donese odluka o prihvaćanju (DA) ili odbacivanju (NE) određene pretpostavke o svojstvima slučajne varijable X. Takva pretpostavka zove se ***statistička hipoteza***, a postupak donošenja odluke o prihvaćanju ili odbacivanju statističke hipoteze zove se ***testiranje***.

Putovi kojima se dolazi do konkretne statističke hipoteze prilikom istraživanja prirodnih i društvenih slučajnih fenomena mogu biti vrlo različiti. Najčešće su to prethodna kvalitativna i kvantitativna promatranja dotične pojave. Iskustvo i intuicija također mogu navesti istraživača da postavi određenu statističku hipotezu o promatranoj pojavi. Znanstvena znatiželja može povući istraživača da pokuša verificirati i nasumce postavljene hipoteze.

Općenito se problem testiranja statističke hipoteze H sastoji u tome da se, na temelju izmjerenih vrijednosti $x_1, \ldots, x_n$ relevantne veličine X, donese odluka o prihvaćanju ili odbacivanju hipoteze H. Budući da se uređena n-torka $(x_1, \ldots, x_n)$ može apstraktno shvatiti kao točka n-dimenzionalnog prostora $\mathbf{R}^n$, riječ je o tome da se skup $\mathbf{R}^n$ podijeli na dva disjunktna dijela C i $\overline{C} = \mathbf{R}^n \backslash C$, pa ako točka $(x_1, \ldots, x_n)$ padne u C, onda će se hipoteza H odbaciti, a ako padne u $\overline{C}$ hipoteza H će se prihvatiti. Skup C zove se ***kritično područje*** hipoteze H.

Prema tome, proces donošenja odluke u pogledu prihvaćanja dane hipoteze H definiran je onda ako je odabrana veličina uzorka n (broj mjerenja) i ako je definirano kritično područje C. Tako definirani postupak zove se ***statistički test*** s fiksiranom veličinom uzorka. Postoje i tzv. ***sekvencijalni testovi*** u kojima nije fiksirana veličina uzorka, već se nakon svakog mjerenja može donijeti odluka DA (prihvaća se hipoteza H), NE (odbacuje se hipoteza H) i DALJE (izvodi se jedno dodatno mjerenje ili više njih).

U statističkim testovima ključnu ulogu ima kritično područje C. Njega treba tako odrediti da sadrži one točke $(x_1, \ldots, x_n) \in \mathbf{R}^n$ u kojima dolazi do ***značajnog*** (***signifikantnog***) ***odstupanja*** od pretpostavljene hipoteze H. Ako rezultati mjerenja $x_1, \ldots, x_n$ upućuju na značajne razlike (recimo, izmjerene količine promatranog sastojka su prevelike) s obzirom na pretpostavljenu hipotezu H (količina ne premašuje kritičnu vrijednost x_0), onda će se hipoteza H odbaciti. Stanovita, ne prevelika, odstupanja od pretpostavljene hipoteze H u nizu mjerenja $x_1, \ldots, x_n$ će se, naravno, tolerirati, jer je osnovna pretpostavka da mjerenja potječu od slučajne varijable X.

Zadatak je, stoga, teorije testiranja statističkih hipoteza da razvije određene metode kojima će se moći razlučiti signifikantna odstupanja od *tolerantnih odstupanja*, te da definira pokazatelje *rizika* za donošenje pogrešne odluke glede hipoteze H.

Da bi se jasnije uočili tipični problemi pri testiranju statističke hipoteze, najprije će se razmotriti jedan jednostavan primjer.

1. primjer

Neka je X relevantna veličina (recimo težina određenog sastojka) za kvalitetu proizvoda, pri čemu je proces proizvodnje takav da se X može smatrati određenom slučajnom varijablom. Proces proizvodnje osigurava zahtijevanu kvalitetu ako je $E[X] = \mu \leq 50$, tj. ako srednja vrijednost težine sastojka u ukupnoj masi proizvoda ne premašuje 50 težinskih jedinica. Na temelju niza od $n = 9$ mjerenja $x_1, \ldots, x_9$ treba testirati hipotezu $H_0 : \mu \leq 50$, tj. donijeti odluku o tome je li zahtijevana kvaliteta postignuta ili nije.

Odmah se vidi da je zapravo riječ o dvije hipoteze, tj. $H_0 : \mu \leq 50$ i $H_1 : \mu > 50$, pri čemu se H_1 zove *alternativna hipoteza* za H_0, koja se pak zove *nul-hipoteza*. To znači da prihvatiti H_0 istodobno znači odbaciti H_1, odnosno odbaciti H_0 znači prihvatiti H_1.

Budući da se rezultati mjerenja $(x_1, \ldots, x_9)$ mogu interpretirati kao točke apstraktnog prostora $\mathbf{R}^9$, kritično područje C za hipotezu H_0 bit će određeni podskup od $\mathbf{R}^9$. Postoji, dakako, beskonačno mnogo mogućnosti za izbor kritičnog područja $C \subset \mathbf{R}^9$, pa se odmah postavlja zadatak da se izbor kritičnog područja, na određeni način vrednuje sa stajališta korektnosti procesa odlučivanja. U promatranom primjeru čini se razumnim kritično područje C definirati tako da ga sačinjavaju one točke iz $\mathbf{R}^9$ u kojima se dobiva prevelika vrijednost aritmetičke sredine $\overline{x} = \frac{1}{9}(x_1 + \ldots + x_9)$. Stoga stavimo

$$C = \{(x_1, \ldots, x_9) \in \mathbf{R}^9 : \overline{x} > 50\}.$$

To, praktički, znači da će se hipoteza H_0 odbacivati kada se, u devet mjerenja, dobije aritmetička sredina veća od 50. Pritom nas je očigledno vodila spoznaja da je aritmetička sredina $\overline{x}$ dobra procjena za nepoznati parametar μ, pa ako je $\overline{x} > 50$, onda se može očekivati da je i nepoznati parametar $\mu > 50$, a to znači da proces proizvodnje ne osigurava zahtijevanu kvalitetu.

Slaba je strana takvog zaključivanja u tome što $\overline{x}$ redovito nije jednako nepoznatom parametru μ, pa se može dobiti, recimo, $\overline{x} = 51$, zbog čega se odbacuje hipoteza H_0, a da je stvarna vrijednost parametra, recimo, $\mu = 49$, što znači da proces proizvodnje osigurava zahtijevanu kvalitetu, odnosno da je hipoteza H_0 stvarno istinita. Stoga se odmah nameće i pitanje kako da se egzktno mjeri rizik odbacivanja istinite hipoteze.

Da bi se odgovorilo na postavljeno pitanje nužno je preciznije definirati matematički model za rješavanje postavljenog zadatka. Može se, recimo, uzeti da je riječ o parametarskom modelu s klasom $\mathcal{P} = \{N(\mu, 36) : \mu \in \mathbf{R}\}$ dopuštenih vjerojatnosnih razdioba. Rezultati mjerenja $(x_1, \ldots, x_9)$ mogu se interpretirati kao vrijednost slučajnog uzorka $(X_1, \ldots, X_9)$, gdje su $X_1, \ldots, X_9$ nezavisne slučajne varijable sa zajedničkom normalnom razdiobom nepoznatog očekivanja μ i poznate varijance $\sigma^2 = 36$. Nadalje se $\overline{x}$ može shvatiti kao vrijednost statistike $\overline{X} = \frac{1}{9}(X_1 + \ldots + X_9)$, za koju se zna da je slučajna varijabla normalne razdiobe $N\left(\mu, \frac{1}{9}\sigma^2\right) = N(\mu, 4)$ (v. (52) u VI.4). To omogućuje da se promatra vjerojatnost odbacivanja hipoteze H_0 u ovisnosti o parametru $t = \mu$, tj. vjerojatnost da slučajna varijabla $\overline{X}$ poprimi vrijednost veću od 50. Pišemo

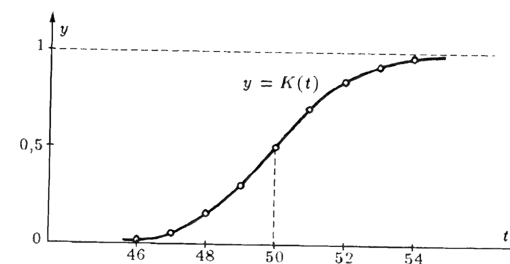
$$P((X_1, \ldots, X_9) \in C) = P(\overline{X} > 50) = 1 - P(\overline{X} \leq 50) = 1 - \Phi\left(\frac{50 - t}{2}\right) =$$
$$= \Phi\left(\frac{t - 50}{2}\right), \quad t \in \mathbf{R},$$

gdje je Φ f.r.v. za standardnu normalnu razdiobu N(0,1). Odmah se vidi da je dobivena određena realna funkcija $t \mapsto K(t)$, $t \in \mathbf{R}$, koja se zove *funkcija snage testa* (*power function*) i čija vrijednost

$$K(t) = \Phi\left(\frac{t - 50}{2}\right) \tag{1}$$

pokazuje vjerojatnost da se hipoteza H_0 odbaci kada je stvarna vrijednost nepoznatog parametra μ jednaka realnom broju t.

Iz formule (1) i slike 13. vidi se, na primjer, da je $K(50) = 0{,}5$, što znači da, u



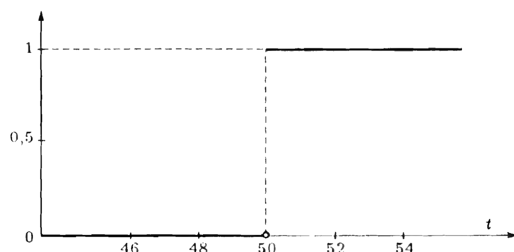
Slika 13. Graf funkcije snage (1)

situaciji kada je stvarna vrijednost parametra $t = 50$, postoji čak 50 % rizika da se odbaci istinita hipoteza.

Mnogima će se činiti da je to preveliki rizik donošenja pogrešne odluke i da treba nešto mijenjati u postupku testiranja, pogotovo zato što je, recimo, za $t = 48$, tj. kada je hipoteza H_0 "debelo" istinita, vjerojatnost odbacivanja hipoteze H_0 jednaka $K(48) = \Phi(-1) \approx 0{,}16 = 16\,\%$. To, praktički, znači da će pri stvarnom stanju u kojem proces proizvodnje osigurava i više od tražene kvalitete, opisani statistički test davati oko 16 % pogrešnih odluka. Rizik donošenja pogrešne odluke past će na manje od 5 %, što se obično smatra tolerantnim rizikom, tek onda ako je $t < 46{,}6$, tj. ako je stvarno očekivanje manje od 46,6.

S druge strane, za $t = 51$, tj. kada je hipoteza H_0 stvarno neistinita, $K(51) = \Phi(0{,}5) \approx 0{,}69 = 69\,\%$, što znači da neistinitu hipotezu odbacujemo sa 69 % vjerojatnosti, a prihvaćamo je sa 31 % vjerojatnosti, što se može smatrati vrlo nepovoljnim.

Sada se može naslutiti kako bi trebala izgledati *idealna funkcija snage* testa u ovom primjeru. Očigledno je, naime, da bi za $t \leq 50$ trebalo biti $K(t) = 0$, a za $t > 50$ trebalo bi biti $K(t) = 1$, tj. vjerojatnost odbacivanja hipoteze H_0 kada je ona istinita trebala bi biti nula, a kada je stvarno neistinita ta vjerojatnost trebala bi biti jedan. Kakva god bila vrijednost nepoznatog parametra μ, primjenom testa s idealnom funkcijom snage rizik donošenja pogrešne odluke bit će nula.



Slika 14. Graf idealne funkcije snage

Ako je, dakle, izabrani matematički model dovoljno vjeran stvarnosti, onda nam teorija statističkog zaključivanja jamči da ćemo, pomoću testa s idealnom funkcijom snage, uvijek donositi korektne odluke. Nevolja je, međutim, u tome što se s konačnim slučajnim uzorkom (konačnim brojem mjerenja) ne može konstruirati statistički test kojemu bi pripadala idealna funkcija snage. No, idealna funkcija snage pokazuje nam čemu treba težiti pri definiranju dobrih testova.

Pokažimo najprije, na već razmotrenome konkretnom primjeru, da se s istom veličinom uzorka $n = 9$ ne može dobiti bitno bolji test. Možemo, recimo, zahtijevati da se kritično područje C' definira tako da rizik odbacivanja hipoteze H_0 ne premaši 5 % ni za koju vrijednost parametra $t \leq 50$. To će se postići tako da se hipoteza H_0 odbacuje kada je vrijednost aritmetičke sredine $\overline{x}$ nešto veća od 50. Koliko ta vrijednost, označimo je sa c, treba iznositi, odredit će se iz uvjeta

(2) $$\max_{t \leq 50}\{P(\overline{X} > c)\} = \max_{t \leq 50} \Phi\left(\frac{t-c}{2}\right) = 0{,}05.$$

Budući da je $t \mapsto \Phi\left(\frac{t-c}{2}\right)$, $t \in \mathbf{R}$, strogo rastuća funkcija, bit će

(3) $$\max_{t \leq 50} \Phi\left(\frac{t-c}{2}\right) = \Phi\left(\frac{50-c}{2}\right),$$

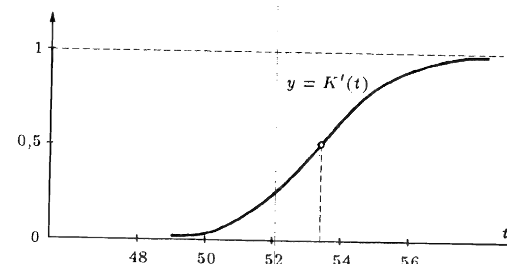
pa iz (2) i (3) odmah proizlazi da je $c \approx 53{,}3$. Definiramo li kritično područje C' kao

$$C' = \{(x_1, \ldots, x_9) \in \mathbf{R}^9 : \overline{x} > 53{,}3\},$$

onda testu pripada funkcija snage

(4) $$K'(t) = \Phi\left(\frac{t-53{,}3}{2}\right), \quad t \in \mathbf{R},$$

čiji graf je skiciran na slici 15.



Slika 15. Graf funkcije snage (4)

Tu se zorno vidi da, kada proces proizvodnje osigurava zahtijevanu kvalitetu, opisani postupak testiranja može, najviše sa 5 % rizika, rezultirati pogrešnom odlukom. Međutim, loša je strana ovog testa što, na primjer, za $t = 51$, tj. kada hipoteza nije istinita, odbacujemo tu hipotezu tek s vjerojatnošću $K'(51) = \Phi(-1{,}15) \approx 0{,}12 = 12\,\%$, tako da je rizik donošenja pogrešne odluke oko 88 %. Iz slike 16. vidi se, dapače, da je za $50 < t \leq 53{,}3$ rizik donošenja pogrešne odluke veći od 50 %.

Testom kojemu pripada kritično područje C' uspjelo se postići da rizik donošenja pogrešne odluke kada je hipoteza H_0 stvarno istinita ($t \leq 50$) ne premašuje 5 %, ali je ostao vrlo širok interval vrijednosti nepoznatog parametra gdje hipoteza H_0 nije istinita, a gdje se H_0 odbacuje s premalom vjerojatnošću, odnosno pogrešna odluka donosi se s prevelikom vjerojatnošću. To je posljedica činjenice da se oblik krivulje u grafu funkcije snage nije promijenio, nego je krivulja na slici 15. translacija krivulje sa slike 13.

Da bi se dobila funkcija snage testa koja će biti sličnija idealnoj funkciji snage, može se postupiti tako da se veličina uzorka n i konstanta c odrede zahtjevima da funkcija snage u točki $t_0 = 50$ poprimi vrijednost 0,05, a u točki $t_1 = 51$ vrijednost 0,95. Pretpostavlja se, naravno, da je odgovarajuće kritično područje C'' opet određeno pomoću vrijednosti statistike $\overline{X}$, tako da se može pisati

$$C'' = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \overline{x} > c\},$$

iz čega proizlazi da pripadna funkcija snage testa K'' ima oblik

(5) $K''(t) = \mathrm{P}((X_1,\ldots,X_n) \in C'') = \mathrm{P}(\overline{X} > c) = 1 - \mathrm{P}(\overline{X} \leq c), \quad t \in \mathbf{R}.$

Statistika $\overline{X} \sim \mathrm{N}\left(t, \frac{36}{n}\right)$, pa se (5) može pisati kao

(6) $$K''(t) = 1 - \Phi\left(\frac{c-t}{6}\sqrt{n}\right) = \Phi\left(\frac{t-c}{6}\sqrt{n}\right), \quad t \in \mathbf{R}.$$

Prvi zahtjev na funkciju snage testa glasi

$$K''(50) = \Phi\left(\frac{51-c}{6}\sqrt{n}\right) = 0{,}05,$$

a drugi zahtjev glasi

$$K''(51) = \Phi\left(\frac{51-c}{6}\sqrt{n}\right) = 0{,}95.$$

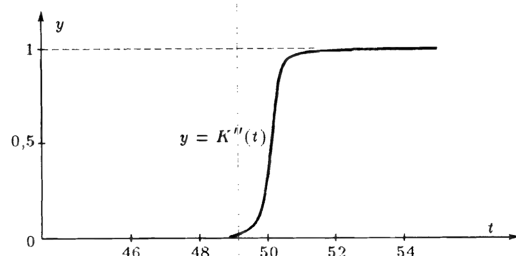
Rješavanjem tog sustava jednadžbi po nepoznanicama c i n dobiva se

$$c \approx 50{,}5, \quad n \approx 392,$$

tako da je odgovarajuća funkcija snage testa

(7) $$K''(t) = \Phi\left(\frac{t-50{,}5}{0{,}3}\right), \quad t \in \mathbf{R}.$$

Njezin graf skiciran je na sl. 16.



Slika 16. Graf funkcije snage (7)

Usporede li se slike 14. i 16. vidi se da je zaista dobivena funkcija snage koja je vrlo slična idealnoj funkciji snage. Za to je, međutim, bio potreban velik broj mjerenja ($n = 392$).

Promatanjem slike 16. razabire se da konstruirani test, s veličinom uzorka $n = 392$ i kritičnim područjem $C'' = \{(x_1,\ldots,x_n) \in \mathbf{R}^n : \overline{x} > 50{,}5\}$, omogućuje donošenje odluke s manje od 5 % rizika za grešku u odlučivanju, za svaki $t < 50$ i svaki $t > 51$. Jedino za $50 \leq t \leq 51$, tj. kada se proces proizvodnje odvija tako da relevantna veličina X ima normalnu razdiobu s očekivanjem između 50 i 51, u konstruiranom će se testu odluke donositi uz rizik greške veći od 5 %.

Prema tome, i uz vrlo veliko n ostaje određeni skup vrijednosti (interval [50,51]) nepoznatog parametra za koji konstruirani test loše funkcionira. Treba, međutim, primijetiti da za sve ostale dopuštene vrijednosti nepoznatog parametra dani test dobro funkcionira.

2. Parametarski test

Kao i u problemu procjene parametra, i ovdje će se pretpostaviti da nepoznati parametar t (može biti i vektorski) pripada zadanom nepraznom skupu Θ dopuštenih vrijednosti i da nepoznata vjerojatnosna razdioba slučajne varijable X pripada zadanoj klasi $\mathcal{P} = \{P_t : t \in \Theta\}$ dopuštenih razdioba vjerojatnosti.

U 1. primjeru hipoteza H_0 odnosila se na parametar $t = \mu$ normalne razdiobe $\mathrm{N}(\mu, 36)$, tako da smo imali skup $\mathbf{R}$ kao skup dopuštenih vrijednosti nepoznatog parametra, i $\mathcal{P} = \{\mathrm{N}(\mu, 36) : \mu \in \mathbf{R}\}$ kao klasu dopuštenih razdioba vjerojatnosti. Hipoteza H_0 opisana je izjavom da nepoznati parametar μ poprima vrijednost koja nije veća od 50, tj. vrijednost iz podskupa $\Theta_0 = \langle -\infty, 50]$ skupa $\Theta = \mathbf{R}$. Prema tome, hipoteza H_0 može se matematički iskazati izrazom $t \in \Theta_0$.

Općenito se u parametarskom testu hipoteza H_0, tzv. ***nul-hipoteza***, izražava tako da se istakne određeni neprazni podskup Θ_0 skupa dopuštenih vrijednosti Θ i zapiše

(8) $$\mathrm{H}_0 : t \in \Theta_0, \quad \Theta_0 \subseteq \Theta,$$

čime se ističe pretpostavka (hipoteza) da promatranoj slučajnoj varijabli X pripada vjerojatnosna razdioba $P_t \in \mathcal{P}$ za koju je $t \in \Theta_0$. Hipoteza

$$\mathrm{H}_1 : t \in \Theta_1, \quad \Theta_1 = \Theta \backslash \Theta_0,$$

zove se ***alternativna hipoteza*** za hipotezu H_0.

Ako je Θ_0 jednočlani skup, tj. $\Theta_0 = \{t_0\}$, onda se govori o ***jednostavnoj hipotezi*** i piše se

$$\mathrm{H}_0 : t = t_0, \quad t_0 \in \Theta.$$

Ako Θ_0 sadrži više od jednog elementa, onda se govori o ***složenoj hipotezi***.

U 1. primjeru hipoteza $\mathrm{H}_0 : \mu \leq 50$ može se zapisati i kao $\mathrm{H}_0 : \mu \in \langle -\infty, 50]$, pa se vidi da je H_0 složena hipoteza.

Označi li se sa C ($C \subseteq \mathbf{R}^n$) kritično područje hipoteze H_0, ***funkcija snage testa*** općenito se definira formulom

(9) $$K(t) = P_t(X_1,\ldots,X_n) \in C), \quad t \in \Theta.$$

$K(t)$ ima, dakle, značenje vjerojatnosti da niz $(x_1,\ldots,x_n)$ mjerenja slučajne varijable X upadne u kritično područje C, odnosno da se odbaci hipoteza H_0, kada je stvarna vrijednost nepoznatog parametra jednaka broju t.

Osim funkcije snage testa, za opisivanje svojstava parametarskog testa upotrebljava se i tzv. ***operativna karakteristika*** testa koja se definira formulom

(10) $$\overline{K}(t) = 1 - K(t) = P_t((X_1, \ldots, X_n) \in \overline{C}), \quad t \in \Theta,$$

gdje je $\overline{C} = \mathbf{R}^n \backslash C$, a $P_t((X_1, \ldots, X_n) \in \overline{C})$ označuje vjerojatnost da niz $(x_1, \ldots, x_n)$ mjerenja slučajne varijable X ne upadne u kritično područje C, kada je stvarna vrijednost nepoznatog parametra jednaka broju t. Vidi se da $\overline{K}(t)$ ima značenje vjerojatnosti da se prihvati hipoteza H_0 kada je t stvarna vrijednost nepoznatog parametra.

Graf operativne karakeristike (*operating characteristic*) zove se ***OC-krivulja*** testa.

Idealna funkcija snage parametarskog testa izražena je formulom

(11a) $$K_0(t) = \begin{cases} 0, & \text{za } t \in \Theta_0 \\ 1, & \text{za } t \in \Theta \backslash \Theta_0, \end{cases}$$

a *idealna operativna karakteristika* formulom

(11b) $$\overline{K}_0(t) = \begin{cases} 1, & \text{za } t \in \Theta_0 \\ 0, & \text{za } t \in \Theta \backslash \Theta_0. \end{cases}$$

Ako funkcija snage parametarskog testa zadovoljava uvjet da je α $(0 \le \alpha \le 1)$ njen maksimum na skupu Θ_0, tj. vrijedi

(12) $$\max_{t \in \Theta_0} K(t) = \alpha,$$

onda se kaže da test ima ***razinu značajnosti*** (***nivo signifikantnosti***) α. Uvjet (12) ekvivalentan je uvjetu

(13) $$\min_{t \in \Theta_0} \overline{K}(t) = 1 - \alpha,$$

pa se može reći da test ima razinu značajnosti α ako vjerojatnost odbacivanja hipoteze H_0, kada je stvarno istinita, ni u kojem slučaju nije veća od α, odnosno da je vjerojatnost prihvaćanja hipoteze H_0, kada je stvarno istinita, bar $1 - \alpha$.

U 1. primjeru imali smo $\alpha = 0{,}5$ za kritično područje C (v. sl. 13), dok je za kritično područje C'' bilo $\alpha = 0{,}05$ (v. sl. 15), kao i za kritično područje C''' (v. sl. 16).

Budući da broj α, na određeni način, utječe na veličinu pripadnoga kritičnog područja C , broj α zove se još i *veličina kritičnog područja.*

Ako je riječ o jednostavnoj hipotezi $H_0 : t = t_0$, očigledno je

(14) $$\max_{t \in \Theta_0} K(t) = K(t_0) = \alpha,$$

odnosno

(15) $$\min_{t \in \Theta_0} \overline{K}(t) = \overline{K}(t_0) = 1 - \alpha,$$

pa se može reći da α označuje vjerojatnost da se odbaci, odnosno $1 - \alpha$ je vjerojatnost da se prihvati istinita jednostavna hipoteza H_0.

2. primjer

Treba konstruirati matematički model za donošenje odluke o hipotezi da se u ljudskoj populaciji rađa jednak broj djevojčica i dječaka. U tu svrhu definirat će se diskretna slučajna varijabla X sa skupom vrijednosti $\{0, 1\}$, pri čemu $X = 0$ označuje rođenje djevojčice, a $X = 1$ rođenje dječaka. Pretpostavlja se, nadalje, da postoji određena vjerojatnost p $(0 < p < 1)$ da se rodi dječak, pa dakle i vjerojatnost $1 - p$ da se rodi djevojčica. Matematički iskazano, to znači da $X \sim \mathrm{B}(1, p)$.

Slučajna pojava spola djeteta pri rođenju matematički je opisana kao slučajna varijabla Bernoullijeve razdiobe $\mathrm{B}(1, p)$ nepoznatog parametra $t = p$, što znači da se usvaja parametarski model s klasom $\mathcal{P} = \{\mathrm{B}(1, p) : p \in \langle 0, 1 \rangle\}$ dopuštenih razdioba vjerojatnosti i skupom $\Theta = \langle 0, 1 \rangle$ dopuštenih vrijednosti nepoznatog parametra. Hipoteza H_0 da se rađa podjednak broj dječaka i djevojčica izrazit će se jednadžbom $p = 0{,}5$. Zadatak se, dakle, sastoji u tome da se, na temelju registriranja spola n novorođenčadi, utvrdi postupak za prihvaćanje, odnosno odbacivanje jednostavne hipoteze $H_0 : p = 0{,}5$.

Zbog slučajnosti promatrane pojave jasno je da će svaki postupak donošenja odluke biti povezan s određenim rizikom za donošenje pogrešne odluke, pa se zadatak može precizirati tako da se unaprijed zahtijeva da rizik odbacivanja hipoteze H_0, kada je stvarno istinita, iznosi $\alpha = 0{,}05$. Može se još reći da se zahtijeva da veličina kritičnog područja C odgovarajućeg testa iznosi $\alpha = 0{,}05$.

Zadatak se, dakle, sveo na to da se, za dane n i α, odredi kritično područje C, pri testiranju jednostavne hipoteze $H_0 : p = 0{,}5$, prema alternativnoj hipotezi $H_1 : p \neq 0{,}5$. Intuicija nas vodi ideji da bi se za rješavanje danog zadatka mogli poslužiti nekim dobrim procjeniteljem nepoznatog parametra p Bernoullijeve razdiobe $\mathrm{B}(1, p)$. U 1. primjeru iz VI.1. vidjeli smo da je aritmetička sredina $\overline{X}$ procjenitelj za p s mnogo dobrih svojstava (nepristranost, konzistentnost, asimptotska normalnost i sl.), pa se može očekivati da, ako je hipoteza H_0 istinita, vrijednost $\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$, tj. relativna frekvencija pojave dječaka među n novorođenčadi neće znatno odstupati od 0,5. Stoga se čini razumnim kritično područje C za hipotezu H_0 definirati tako da ga sačinjavaju one točke $(x_1, \ldots, x_n) \in \mathbf{R}^n$ za koje se dobiva $\overline{x}$ "previše" udaljeno od 0,5. Stavimo zato

(16) $$C = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : |\overline{x} - 0{,}5| > c\},$$

i pokušajmo odrediti $c \ge 0$ tako da kritično područje C ima veličinu $\alpha = 0{,}05$.

Kada bismo poznavali c, onda bi, prema (10), operativna karakteristika odgovarajućeg testa izgledala

(17) $$\overline{K}(t) = P_t((X_1, \ldots, X_n) \in \overline{C}) = P_t(|\overline{X} - 0{,}5| \le c), \quad t \in \langle 0, 1 \rangle.$$

Za svako $n \in \mathbf{N}$ procjenitelju $\overline{X}$ pripada diskretna razdioba vjerojatnosti opisana formulom (4) u VI.1, a za velike n može se približno uzeti da $\overline{X} \sim \mathrm{N}(t, \frac{1}{n}(1-t))$, jer

je $\overline{X}$ asimptotski normalan procjenitelj nepoznatog parametra $t = p$. To omogućuje da se (17) zapiše u obliku

$$(18)\qquad \overline{K}(t) = P_t(0{,}5 - c \leq \overline{X} \leq 0{,}5 + c) \approx$$
$$\approx \Phi\left(\frac{0{,}5 + c - t}{\sqrt{t(1-t)}}\sqrt{n}\right) - \Phi\left(\frac{0{,}5 - c - t}{\sqrt{t(1-t)}}\sqrt{n}\right), \quad t \in \langle 0, 1\rangle.$$

Budući da je $H_0 : t = 0{,}5$ jednostavna hipoteza, u skladu sa (15), zaključuje se da mora vrijediti

$$(19)\qquad \overline{K}(t_0) = \overline{K}(0{,}5) = 2\Phi(2c\sqrt{n}) - 1 = 1 - \alpha,$$

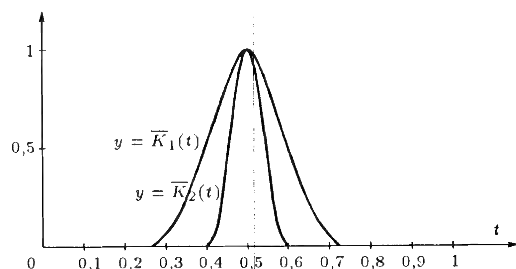
iz čega se dobiva

$$(20)\qquad c = \frac{1}{2\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Posebno, za $\alpha = 0{,}05$ i $n = 100$, proizlazi $c \approx 0{,}1$, što znači da, promatrajući spol stotinu novorođenčadi, hipotezu da se u ljudskoj populaciji rađa podjednaki broj djevojčica i dječaka odbacujemo onda ako se dobije $|\overline{x} - 0{,}5| > 0{,}1$, tj. ako se nađe više od 60 dječaka (djevojčica), pri čemu se, dakako, s vjerojatnošću $\alpha = 0{,}05$ može odbaciti i istinita hipoteza.

Za $n = 100$ i $c = 0{,}1$ formula (18) postaje

$$(21)\qquad \overline{K}_1(t) = \Phi\left(\frac{0{,}6 - t}{\sqrt{t(1-t)}} \cdot 10\right) - \Phi\left(\frac{0{,}4 - t}{\sqrt{t(1-t)}} \cdot 10\right), \quad t \in \langle 0, 1\rangle.$$



Slika 17. Graf funkcije (21) i (22)

Iz pripadne OC-krivulje, prikazane na slici 17, vidi se da konstruirani test osigurava prihvaćanje hipoteze H_0 s velikom vjerojatnošću $1 - \alpha = 0{,}95$ kada je stvarno istinita. Međutim, vidi se i to da se hipoteza H_0 prihvaća s vrlo velikom vjerojatnošću i za neke vrijednosti parametra $t = p$ kada stvarno nije istinita. Tako se, na primjer, za $t = 0{,}4$ prihvaća s vjerojatnošću 0,5, a isto tako i za $t = 0{,}6$. Tek za $t < 0{,}3$ i $t > 0{,}7$ vjerojatnost prihvaćanja neistinite hipoteze H_0 postaje vrlo bliska nuli.

Želimo li konstruirati test koji će imati operativnu karakteristiku sličniju idealnoj, moramo uzeti veći uzorak. Za $n = 625$, na primjer, dobiva se, prema (20), $c \approx 0{,}04$ i pripadna operativna karakteristika glasi

$$(22)\qquad \overline{K}_2(t) = \Phi\left(\frac{0{,}54 - t}{\sqrt{t(1-t)}} \cdot 25\right) - \Phi\left(\frac{0{,}46 - t}{\sqrt{t(1-t)}} \cdot 25\right), \quad t \in \langle 0, 1\rangle.$$

Ona ima svojstvo da je $\overline{K}_2(0{,}4) = \overline{K}_2(0{,}6) \approx 0$, a njen graf je također prikazan na slici 17. Taj test, osim što osigurava da se s velikom vjerojatnošću $1 - \alpha = 0{,}95$ prihvaća hipoteza H_0 kada je stvarno istinita, osigurava još i to da se hipoteza H_0 prihvaća s vrlo malom vjerojatnošću, praktički jednakom nuli, za svaku vrijednost parametra t izvan uskog intervala $[0{,}5 - \delta; 0{,}5 + \delta]$ ($\delta \approx 0{,}04$), tj. kada H_0 stvarno nije istinita.

Oba razmotrena primjera (1. i 2. primjer) upućuju na pojavu dvaju tipova pogrešaka pri testiranju parametarskih statističkih hipoteza. Odbacivanje nul-hipoteze $H_0 : t \in \Theta_0$, kada je stvarno istinita, zove se ***pogreška prve vrste***, dok se prihvaćanje nul-hipoteze H_0, kada je stvarno neistinita, zove ***pogreška druge vrste***.

Funkcija snage, odnosno operativna karakteristika testa omogućuje da se uoči vjerojatnost pogreške prve i druge vrste. Već je rečeno da se s konačnom veličinom uzorka n ne može dobiti idealna funkcija snage testa, pa se za konačno fiksirano n obično postavlja ovakav zadatak: Konstruirati takav test, tj. odrediti pripadno kritično područje C tako da razina značajnosti testa, što je ujedno i najveća moguća pogreška prve vrste, iznosi zadani broj α ($0 < \alpha < 1$) i da istodobno najveća moguća vrijednost pogreške druge vrste bude minimalna.

Ako bi se uspio konstruirati takav test, onda bi se time dobio najbolji mogući test za dano α i n. Međutim, za mnoge probleme testiranja statističkih hipoteza takav test ne postoji. Stoga je razumljivo da se najprije razmotri najjednostavniji slučaj problema testiranja parametarske statističke hipoteze u kojem se pretpostavlja da je skup Θ dopuštenih vrijednosti nepoznatog parametra t dvočlani skup $\Theta = \{t_0, t_1\}$. To znači da je nul-hipoteza jednostavna hipoteza $H_0 : t = t_0$, dok je alternativna hipoteza također jednostavna hipoteza $H_1 : t = t_1$. Imajući na umu (9) i (14), uvjeti za najbolji test, odnosno za odgovarajuće kritično područje C_0, u opisanom jednostavnom slučaju glase

$$(23)\qquad P_0((X_1, \ldots, X_n) \in C_0) = \alpha,$$

$$(24)\qquad P_1((X_1, \ldots, X_n) \in C_0) \geq P_1((X_1, \ldots, X_n) \in C), \quad \forall C \subseteq \mathbf{R}^n,$$

gdje $P_0 = P_{t_0}$ označuje vjerojatnost pri istinitosti hipoteze H_0, $P_1 = P_{t_1}$ označuje vjerojatnost pri istinitosti hipoteze H_1, dok su C_0 i C kritična područje za hipotezu H_0 veličine α.

Iz (23) i (24) razabire se da se problem nalaženja najboljeg testa sastoji u tome da se odredi ono kritično područje C_0, veličine α, za koje je vjerojatnost "upadanja" niza podataka $(x_1, \ldots, x_n)$ u C_0, kada je stvarna vjerojatnosna razdioba P_1, veća ili jednaka od vjerojatnosti "upadanja" toga niza u bilo koji drugi podskup $C \subseteq \mathbf{R}^n$, pri čemu vrijedi $P_0((X_1, \ldots, X_n) \in C) = \alpha$.

3. Neyman-Pearsonova lema

Rekli smo već da definirati test, uz zadanu veličinu uzorka $n \in \mathbf{N}$, znači definirati pripadno kritično područje C kao određeni podskup od $\mathbf{R}^n$. Odabere li se kritično područje $C_0 \subseteq \mathbf{R}^n$ nul-hipoteze H_0 tako da vrijedi (23) i (24), onda se govori o *najboljem kritičnom području* veličine α za testiranje jednostavne hipoteze $\mathrm{H}_0 : t = t_0$, prema alternativnoj jednostavnoj hipotezi $\mathrm{H}_1 : t = t_1$.

Da bi se bolje shvatio smisao Neyman-Pearsonove leme, kojom se rješava problem nalaženja najboljega kritičnog područja pri testiranju jednostavne hipoteze H_0, prema alternativnoj jednostavnoj hipotezi H_1, najprije će se razmotriti jedan primjer.

3. primjer

Zamislimo da imamo dva novčića, od kojih je jedan pravilan tj. vjerojatnost pojavljivanja grba iznosi $p_0 = 0{,}5$, a drugi je nepravilan i kod njega je vjerojatnost pojavljivanja grba $p_1 = 0{,}4$. Pretpostavlja se da se pravilnost novčića ne može prepoznati po njegovim vanjskim osobinama, pa će se odluka o njegovoj pravilnosti donijeti na temelju bacanja novčića, recimo, 10 puta i registriranja broja grbova X. Očigledno je X slučajna varijabla binomne razdiobe $\mathrm{B}(10, p)$, gdje je p nepoznati parametar, za koji se zna da može poprimiti vrijednosti $p_0 = 0{,}5$ i $p_1 = 0{,}4$. Zadatak se, stoga, može formulirati kao testiranje jednostavne hipoteze $\mathrm{H}_0 : p = 0{,}5$, prema alternativnoj jednostavnoj hipotezi $\mathrm{H}_1 : p = 0{,}4$.

U tabl. 1 navedene su vjerojatnosti da diskretna slučajna varijabla X poprimi odgovarajuće vrijednosti iz skupa $A = \{0, 1, \ldots, 10\}$ mogućih vrijednosti, i to najprije uz pretpostavku da je $p = 0{,}5$, a zatim uz pretpostavku da je $p = 0{,}4$.

Tablica 1.

x	0	1	2	3	4	5	6	7	8	9	10
$P_0(X = x)$	0,001	0,010	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,010	0,001
$P_1(X = x)$	0,006	0,040	0,121	0,215	0,251	0,201	0,111	0,042	0,011	0,002	0,0001
$\dfrac{P_0(X = x)}{P_1(X = x)}$	0,17	0,25	0,36	0,54	0,82	1,22	1,85	2,78	4,00	5,00	10,00

Uzmimo $n = 1$, tj. na temelju jednog mjerenja slučajne varijable X pokušajmo naći najbolje kritično područje $C_0 \subseteq \mathbf{R}$, veličine $\alpha = 0{,}055 = 5{,}5\,\%$. Budući da je

$$\begin{aligned} &P_0(X = 0) + P_0(X = 1) + P_0(X = 2) = \\ &= P_0(X = 8) + P_0(X = 9) + P_0(X = 10) = \alpha = 5{,}5\,\%, \end{aligned}$$

može se reći da kritična područja $C_0 = \{x \in \mathbf{R} : x \leq 2\}$ i $C = \{x \in \mathbf{R} : x \geq 8\}$, hipoteze H_0, imaju veličinu $\alpha = 0{,}055$, jer očigledno vrijedi $P_0(X \in C_0) = P_0(X \in C) = \alpha$. Iz tabl. 1 također se vidi da je

$$(25) \quad P_1(X \in C_0) = P_1(X = 0) + P_1(X = 1) + P_1(X = 2) = 0{,}167 = 16{,}7\,\%,$$

dok je

$$(26) \quad P_1(X \in C) = P_1(X = 8) + P_1(X = 9) + P_1(X = 10) = 0{,}013 = 1{,}3\%.$$

Iz (25) i (26) proizlazi da je $P_1(X \in C_0) > P_1(X \in C)$, pa se može zaključiti da je C_0 bolje kritično područje veličine α za testiranje hipoteze $\mathrm{H}_0 : p = 0{,}5$, prema alternativnoj hipotezi $\mathrm{H}_1 : p = 0{,}4$, nego što je kritično područje C. Budući da su C_0 i C jedina kritična područja zadane veličine α, slijedi da je C_0 najbolje kritično područje veličine α.

Prema tome, najbolji test sa $n = 1$ i $\alpha = 0{,}055$ za prepoznavanje novčića funkcionira tako da se hipoteza H_0 (novčić je pravilan) odbacuje onda ako se prilikom 10 bacanja toga novčića dobije manje od 3 grba. U protivnom, hipoteza H_0 se prihvaća. Pogreška prve vrste, tj. vjerojatnost da se odbaci istinita hipoteza, iznosi $\alpha = 5{,}5\,\%$, a pogreška druge vrste, tj. vjerojatnost da se prihvati neistinita hipoteza, iznosi $\beta_0 = 1 - P_1(X \in C_0) = 83{,}3\,\%$. Pogreška prve vrste je snošljivo velika, dok je pogreška druge vrste očigledno prevelika, ali to je najbolje što se može postići s veličinom uzorka $n = 1$ (jednokratnim mjerenjem slučajne varijable X). Kada bi se načinio test s kritičnim područjem C, za koje je pogreška prve vrste također $\alpha = 5{,}5\,\%$, pripadna pogreška druge vrste bila bi još veća ($\beta = 1 - P_1(X \in C) = 98{,}7\,\%$).

Ovaj primjer pokazuje, a to će vrijediti i općenito, da se najbolje kritično područje C_0 sastoji od onih točaka $(x_1, \ldots, x_n)$ prostora $\mathbf{R}^n$, za koje vrijedi da je $P_0(X_1 = x_1, \ldots, X_n = x_n)$ mnogo manje od $P_1(X_1 = x_1, \ldots, X_n = x_n)$, odnosno gdje je omjer $\dfrac{P_0(X_1 = x_1, \ldots, X_n = x_n)}{P_1(X_1 = x_1, \ldots, X_n = x_n)}$ dovoljno malen.

U danom je primjeru $n = 1$ i u tabl. 1. navedene su vrijednosti odgovarajućih omjera, gdje se vidi da se za $x_1 = x = 0$ dobiva vrijednost 0,17, za $x = 1$ dobiva se 0,25, a za $x = 2$ dobiva se vrijednost 0,36 toga omjera. Prirodno se, stoga, nameće ideja da se najbolje kritično područje C_0 općenito odredi iz zahtjeva da promatrani omjer bude manji od unaprijed zadanoga pozitivnog broja c. O tome upravo govori *Neyman-Pearsonova lema*:

> Neka su $n \in \mathbf{N}$ i α $(0 < \alpha < 1)$ zadani brojevi i $\mathrm{H}_0 : t = t_0$ nul-hipoteza, uz alternativnu hipotezu $\mathrm{H}_1 : t = t_1$, koje se odnose na diskretnu ili kontinuiranu slučajnu varijablu X. Neka je $t \mapsto \boldsymbol{L}(t)$, $t \in \Theta = \{t_1, t_2\}$, pripadna funkcija vjerodostojnosti (v. (36a) i (36b) u VI.3) i neka je c takav pozitivan broj da su zadovoljeni uvjeti
>
> (i) $$P_0((X_1, \ldots, X_n) \in C_0) = \alpha,$$
>
> (ii) $$\frac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)} \leq c, \quad za \quad (x_1, \ldots, x_n) \in C_0,$$
>
> (iii) $$\frac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)} \geq c, \quad za \quad (x_1, \ldots, x_n) \in \overline{C}_0 = \mathbf{R}^n \backslash C_0.$$
>
> Tada je C_0 najbolje kritično područje veličine α za testiranje jednostavne hipoteze H_0, prema alternativnoj jednostavnoj hipotezi H_1.

Dokaz Neyman-Pearsonove leme za diskretnu slučajnu varijablu X analogan je dokazu za kontinuiranu slučajnu varijablu X, samo što se umjesto integrala pojavljuju sume. Stoga će se provesti samo dokaz za kontinuirani sučaj, gdje se pretpostavlja da slučajnoj varijabli X pripada f.g.v. $x \mapsto f_t(x)$, $x \in \mathbf{R}$, tako da pripadna funkcija vjerodostojnosti glasi

(27) $$\boldsymbol{L}(t) = f_t(x_1) \cdot \ldots \cdot f_t(x_n), \quad (x_1, \ldots, x_n) \in \mathbf{R}^n, \quad t \in \Theta.$$

Budući da vrijednost $\boldsymbol{L}(t)$ ima značenje gustoće vjerojatnosti slučajnog vektora $(X_1, \ldots, X_n)$, kada nepoznati parametar ima vrijednost t, za svaki $C \subseteq \mathbf{R}^n$ može se pisati

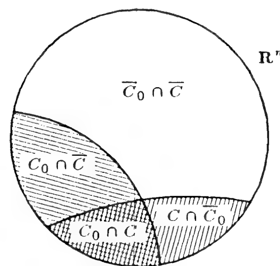
(28) $$P_t((X_1, \ldots, X_n) \in C) = \int_C \boldsymbol{L}(t)\, \mathrm{d}\boldsymbol{x},$$

gdje znak $\int_C$ označuje n-struki integral po skupu C, dok je $\mathrm{d}\boldsymbol{x} = \mathrm{d}x_1 \cdot \ldots \cdot \mathrm{d}x_n$.

Ako je $C_0 \subseteq \mathbf{R}^n$ jedino kritično područje veličine α, onda je lema očigledno istinita. Ako, pak, postoji kritično područje C veličine α i $C \neq C_0$, onda vrijedi

(29) $$C_0 = (C_0 \cap C) \cup (C_0 \cap \overline{C}),$$

(30) $$C = (C_0 \cap C) \cup (C \cap \overline{C}_0).$$



Slika 18. Skica odnosa skupova C i C_0

Budući da su $C_0 \cap C$ i $C_0 \cap \overline{C}$, kao i $C_0 \cap C$ i $C \cap \overline{C}_0$ disjunktni skupovi (v. sl. 18), iz (29), (30) i svojstva aditivnosti integrala proizlazi

$$\int_{C_0} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} = \int_{C_0 \cap C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} + \int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x},$$

$$\int_{C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} = \int_{C_0 \cap C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} + \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x},$$

tako da se može pisati

(31) $$\int_{C_0} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} - \int_{C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} = \int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} - \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x}.$$

Iz uvjeta (ii) proizlazi da je $\boldsymbol{L}(t_0) \leq c\boldsymbol{L}(t_1)$ za svaki $\boldsymbol{x} = (x_1, \ldots, x_n) \in C_0$, a pogotovo za $\boldsymbol{x} \in C_0 \cap \overline{C}$, pa na temelju svojstva monotonosti integrala slijedi

(32) $$\int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} \geq \frac{1}{c} \int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x}.$$

Iz uvjeta (iii), pak proizlazi da je $\boldsymbol{L}(t_0) \geq c\boldsymbol{L}(t_1)$ za svaki $\boldsymbol{x} \in \overline{C}_0$, pa stoga i za $\boldsymbol{x} \in \overline{C}_0 \cap C$, iz čega proizlazi

(33) $$\int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} \leq \frac{1}{c} \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x}.$$

Iz (32) i (33) dobiva se

$$\int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} - \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} \geq \frac{1}{c} \left[\int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} - \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} \right],$$

što zajedno sa (31) omogućuje zaključivanje da je

$$\int_{C_0} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} - \int_{C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} \geq \frac{1}{c} \left[\int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} - \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} \right] =$$

$$= \frac{1}{c} \left[\int_{C_0 \cap \overline{C}} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} + \int_{C_0 \cap C} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} - \int_{C_0 \cap C} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} - \int_{\overline{C}_0 \cap C} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} \right] =$$

$$= \frac{1}{c} \left[\int_{C_0} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} - \int_{C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} \right].$$

Imajući na umu uvjet (i) i (28) odmah se vidi da je $\int_{C_0} \boldsymbol{L}(t_0)\, \mathrm{d}\boldsymbol{x} = \alpha$, a budući da smo pretpostavili da i kritično područje C ima veličinu α, onda je i $\int_C \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} = \alpha$, tako da se konačno dobiva

$$\int_{C_0} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} - \int_{C} \boldsymbol{L}(t_1)\, \mathrm{d}\boldsymbol{x} \geq 0,$$

što zapisano prema (28) daje

$$P_1((X_1, \ldots, X_n) \in C_0) \geq P_1((X_1, \ldots, X_n) \in C).$$

Prema (24) to upravo znači da je C_0 najbolje kritično područje veličine α za testiranje hipoteze $H_0 : t = t_0$, prema alternativnoj hipotezi $H_1 : t = t_1$, što i izriče Neyman-Pearsonova lema.

Primjedba

Neyman-Pearsonova lema ističe relacije (i), (ii) i (iii) kao dovoljan uvjet da C_0 bude najbolje kritično područje. Može se, međutim, dokazati (v. [25]) da su to i nužni uvjeti.

4. primjer

Primjenom Neyman-Pearsonove leme treba konstruirati najbolji test, sa zadanom veličinom uzorka n i razinom značajnosti α, za testiranje jednostavne hipoteze $H_0 : \mu = \mu_0$, prema jednostavnoj alternativnoj hipotezi $H_1 : \mu = \mu_1$, pri čemu se pretpostavlja da je vrijednost slučajnog uzorka $(x_1, \ldots, x_n)$ dobivena mjerenjem slučajne varijable $X \sim N(\mu, 1)$.

Stavimo $t = \mu$, pa se pripadna f.g.v. može izraziti formulom

$$f_t(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-t)^2\right], \tag{34}$$

iz čega, na temelju (27), slijedi da pripadna funkcija vjerodostojnosti glasi

$$\boldsymbol{L}(t) = (2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-t)^2\right]. \tag{35}$$

Stavljajući u (35) $t = \mu_0$ i zatim $t = \mu_1$, nalazi se da je

$$\frac{\boldsymbol{L}(\mu_0)}{\boldsymbol{L}(\mu_1)} = \frac{\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right]}{\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right]} = \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\mu_0)^2-(x_i-\mu_1)^2\right]\right) =$$

$$= \exp\left[(\mu_0-\mu_1)\sum_{i=1}^{n}x_i - \frac{n}{2}(\mu_0^2-\mu_1^2)\right].$$

Budući da je $\sum_{i=1}^{n} x_i = n\overline{x}$, uvjet (ii) iz Neyman-Pearsonove leme glasi

$$\exp\left[n\overline{x}(\mu_0-\mu_1) - \frac{n}{2}(\mu_0^2-\mu_1^2)\right] \le c,$$

odnosno, nakon logaritmiranja glasi

$$\overline{x}(\mu_0-\mu_1) \le \frac{1}{n}\ln c + \frac{1}{2}(\mu_0^2-\mu_1^2). \tag{36}$$

Daljnji zaključci ovise o tome je li $\mu_0 < \mu_1$ ili je $\mu_0 > \mu_1$. Ako je $\mu_0 > \mu_1$, tj. $\mu_0 - \mu_1 > 0$, onda (36) postaje

$$\overline{x} \le \frac{\ln c}{n(\mu_0-\mu_1)} + \frac{1}{2}(\mu_0+\mu_1) = c_0, \tag{37}$$

tako da najbolje kritično područje C_0 ima oblik

$$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \overline{x} \le c_0\}. \tag{38}$$

Iz uvjeta (i), da C_0 ima veličinu α, proizlazi relacija

$$P_0(\overline{X} \le c) = \alpha. \tag{39}$$

Budući da $\overline{X} \sim N\left(\mu, \frac{1}{n}\right)$, (39) se može pisati kao

$$\Phi[\sqrt{n}(c_0-\mu_0)] = \alpha,$$

iz čega proizlazi

$$c_0 = \mu_0 + \frac{1}{\sqrt{n}}\Phi^{-1}(\alpha). \tag{40}$$

Sa (38) i (40) određeno je najbolje kritično područje C_0 za testiranje hipoteze $H_0 : \mu = \mu_0$, prema alternativnoj hipotezi $H_1 : \mu = \mu_1$, za slučaj $\mu_1 < \mu_0$. Ako je $\alpha < 0{,}5$, a uobičajene vrijednosti za α su 0,01 i 0,05, tada je $\Phi^{-1}(\alpha) < 0$, pa iz (40) slijedi da je $c_0 < \mu_0$, a to znači da se hipoteza H_0 odbacuje onda ako se mjerenjem dobije premalena vrijednost uzoračke aritmetičke sredine $\overline{x}$, u usporedbi s pretpostavljenom vrijednosti μ_0, jer je tada "logičnije" prihvatiti hipotezu H_1.

Uzme li se, na primjer, $n = 9$ i $\alpha = 0{,}05$, iz (40) se dobiva $c_0 = \mu_0 - 0{,}55$, tako da će, za svaki $\mu_1 < \mu_0$, test s kritičnim područjem

$$C_0 = \{(x_1, \ldots, x_9) \in \mathbf{R}^9 : \overline{x} \le \mu_0 - 0{,}55\}$$

biti najbolji test za testiranje hipoteze $H_0 : \mu = \mu_0$, prema alternativnoj hipotezi $H_1 : \mu = \mu_1$.

Vjerojatnost pogreške prve vrste općenito se podudara s razinom značajnosti α, dok vjerojatnost pogreške druge vrste, tj. vjerojatnost da se prihvati hipoteza H_0 kada stvarno nije istinita, jest

$$\begin{aligned}\beta_0 &= 1 - P_1((X_1, \ldots, X_n) \in C_0) = 1 - P_1(\overline{X} \le c_0) = \\ &= 1 - \Phi[\sqrt{n}(\mu_0-\mu_1) + \Phi^{-1}(\alpha)].\end{aligned} \tag{41}$$

Tako se, na primjer, za $n = 9$, $\alpha = 0{,}05$ i $\mu_0 - \mu_1 = 1$ dobiva

$$\beta_0 = 1 - \Phi(3 - 1{,}65) = 1 - \Phi(1{,}35) \approx 0{,}08.$$

S uzorkom veličine $n = 100$ dobili bismo $\beta_0 \approx 0$, pa se može reći da se opisanim testom s visokom pouzdanošću ($\alpha = 0{,}05, \beta_0 \approx 0$) razlučuju dvije normalne razdiobe iste varijance ($\sigma^2 = 1$), čija se očekivanja razlikuju za jedinicu ($\mu_0 - \mu_1 = 1$).

Iz (41) se razabire da se vjerojatnost pogreške druge vrste smanjuje s povećanjem uzorka, a također i s povećanjem razlike očekivanja $(\mu_0 - \mu_1)$ pretpostavljenih normalnih razdioba. No, i intuicija nas upućuje na to da će se lakše razlučiti, na temelju n mjerenja, normalne razdiobe čija očekivanja se više razlikuju, tako da formula (41) daje kvantitativnu mjeru za spomenutu intuitivnu spoznaju.

U slučaju $\mu_1 > \mu_0$, nejednakost (36) dijeli se negativnim brojem $\mu_0 - \mu_1$ i stoga postaje

$$\overline{x} \geq \frac{\ln c}{n(\mu_0 - \mu_1)} + \frac{1}{2}(\mu_0 + \mu_1) = c_0',$$

pa najbolje kritično područje C_0' ima oblik

(42) $$C_0' = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \overline{x} \geq c_0'\}.$$

Uvjet (i) dovodi do jednadžbe

$$P_0(\overline{X} \geq c_0') = \alpha,$$

iz koje odmah proizlazi

$$1 - \Phi[(c_0' - \mu_0)\sqrt{n}] = \alpha,$$

i dalje

(43) $$c_0' = \mu_0 + \frac{1}{\sqrt{n}}\,\Phi^{-1}(1-\alpha).$$

Ako je $\alpha < 0{,}5$, onda je $\Phi^{-1}(1-\alpha) = -\Phi^{-1}(\alpha) > 0$, pa se iz (42) i (43) razabire da će se hipoteza H_0 odbaciti onda ako se dobije prevelika vrijednost $\overline{x}$, u odnosu na pretpostavljenu vrijednost μ_0. Tada je očigledno razumnije prihvatiti hipotezu H_1, koja pretpostavlja veću vrijednost μ_1 nepoznatog očekivanja.

Iz (40), kao i iz (43), vidi se da c_0 (c_0') ne ovisi o μ_1, što znači da će test s kritičnim područjem C_0 (C_0') biti najbolji test za testiranje jednostavne hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \mu = \mu_1$, za svaki $\mu_1 < \mu_0$ $(\mu_1 > \mu_0)$.

4. Jednoliko najsnažniji test

Razmatranja u 4. primjeru upućuju nas na ideju kako da se općenito definira najbolji test za testiranje jednostavne hipoteze $\mathrm{H}_0 : t = t_0$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : t \in \Theta_1$, $\Theta_1 = \Theta \backslash \{t_0\}$.

Ako je C_0 najbolje kritično područje veličine α, za testiranje jednostavne hipoteze $\mathrm{H}_0 : t = t_0$, prema jednostavnoj alternativnoj hipotezi $\mathrm{H}_1^1 : t = t_1$, za svaki $t_1 \in \Theta_1$, onda se test s kritičnim područjem C_0 zove ***jednoliko najsnažniji test*** za testiranje jednostavne hipoteze $\mathrm{H}_0 : t = t_0$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : t \in \Theta_1$.

Odmah se vidi da je test iz 4. primjera s kritičnim područjem C_0, definiranim u (38), jednoliko najsnažniji test za testiranje jednostavne hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \mu < \mu_0$. Isto tako je test s kritičnim područjem C_0', definiranim u (42), jednoliko najsnažniji test za testiranje jednostavne hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \mu > \mu_0$.

To bi nas moglo navesti na pomisao da se može konstruirati jednoliko najsnažniji test za testiranje jednostavne hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \mu \neq \mu_0$, tako da se za kritično područje C uzme skup

(44) $$C = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : |\overline{x} - \mu_0| \geq \delta\},$$

gdje je $\delta = -\frac{1}{\sqrt{n}}\,\Phi^{-1}\left(\frac{\alpha}{2}\right)$ određen tako da C ima veličinu α, tj. da test ima razinu značajnosti α. Međutim, taj test nije jednoliko najsnažniji test, jer kritično područje C, definirano u (44), očigledno nije najbolje kritično područje za testiranje jednostavne hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema jednostavnoj alternativnoj hipotezi $\mathrm{H}_1 : \mu = \mu_1$ $(\mu_1 \neq \mu_0)$. Vidjeli smo, naime, da je za $\mu_1 < \mu_0$ najbolje kritično područje $C_0 \neq C$, dok je za $\mu_1 > \mu_0$ najbolje kritično područje $C_0' \neq C$. Štoviše, vidi se da ne može egzistirati jadnoliko najsnažniji test za testiranje hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \mu \neq \mu_0$, jer ne postoji jedinstveno kritično područje koje bi bilo najbolje, uz alternativnu hipotezu $H_1' : \mu = \mu_1$, za slučaj $\mu_1 < \mu_0$ i također za slučaj $\mu_1 > \mu_0$.

Iz definicije jednoliko najsnažnijeg testa za testiranje jednostavne hipoteze $\mathrm{H}_0 : t = t_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : t \in \Theta_1$ $(\Theta_1 = \Theta \backslash \{t_0\})$, proizlazi da funkcija snage K_0 jednoliko najsnažnijeg testa zadovoljava uvjete

(45) $$K_0(t_0) = \alpha,$$

(46) $$K_0(t) \geq K(t), \quad \forall t \in \Theta_1,$$

gdje je K funkcija snage bilo kojega drugog testa sa zadanom veličinom uzorka n i veličinom kritičnog područja α.

Budući da je $\beta(t) = 1 - K(t)$, $t \in \Theta_1$, vjerojatnost pogreške druge vrste, iz (46) slijedi da je

(47) $$\beta_0(t) = 1 - K_0(t) \leq 1 - K(t) = \beta(t), \quad t \in \Theta_1,$$

što pokazuje da u jednoliko najsnažnijem testu vjerojatnost pogreške druge vrste nije veća od vjerojatnosti pogreške druge vrste u bilo kojem drugom testu s jednakim n i α.

Tako, na primjer, funkcija snage K_0 jednoliko najsnažnijeg testa za testiranje hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \mu < \mu_0$, kojemu pripada kritično područje C_0, definirano u (38), glasi

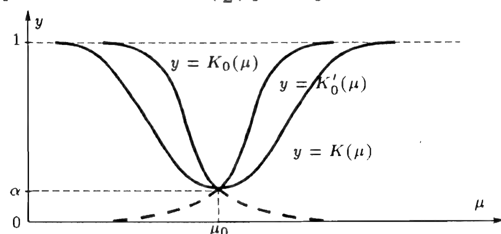
(48) $$K_0(\mu) = P_\mu(\overline{X} \leq c_0) = \Phi[(\mu_0 - \mu)\sqrt{n} + \Phi^{-1}(\alpha)], \quad \mu \in \langle -\infty, \mu_0].$$

Funkcija snage, K_0' jednoliko najsnažnijeg testa za testiranje hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \mu > \mu_0$, kojemu pripada kritično područje C_0', definirano u (42), glasi

(49) $$K_0'(\mu) = P_\mu(\overline{X} \geq c_0) = \Phi[(\mu - \mu_0)\sqrt{n} + \Phi^{-1}(\alpha)], \quad \mu \in [\mu_0, \infty\rangle,$$

dok funkcija snage K testa za testiranje hipoteze $H_0 : \mu = \mu_0$, prema alternativnoj hipotezi $H_1 : \mu \neq \mu_0$, s kritičnim područjem C, definiranim u (44), glasi

$$(50)\quad K(\mu) = \Phi\Big[(\mu_0 - \mu)\sqrt{n} + \Phi^{-1}\Big(\frac{\alpha}{2}\Big)\Big] + \Phi\Big[(\mu - \mu_0)\sqrt{n} + \Phi^{-1}\Big(\frac{\alpha}{2}\Big)\Big], \quad \mu \in \mathbf{R}.$$



Slika 19. Skica grafova funkcija K_0, K_0' i K

U prvom i drugom slučaju imamo jednoliko najsnažnije testove, što znači da ne postoji test čija bi funkcija snage imala graf iznad odgovarajućeg grafa na slici 19. U trećem, pak, slučaju, gdje nam se intuitivno čini da smo odabrali najbolje kritično područje C za testiranje jednostavne hipoteze $H_0 : \mu = \mu_0$, prema složenoj alternativnoj hipotezi $H_1 : \mu \neq \mu_0$, dobiva se graf funkcije snage koji je svagdje, osim u točki μ_0, ispod grafa funkcije K_0, odnosno K_0'. Testira li se jednostavna hipoteza $H_0 : \mu = \mu_0$, prema jednostavnoj alternativnoj hipotezi $H_1 : \mu = \mu_1$, za svaki $\mu_1 \neq \mu_0$ postoji snažniji test, tj. bolje kritično područje od onoga, definiranog u (44), koje se čini najboljim za testiranje jednostavne hipoteze $H_0 : \mu = \mu_0$, prema složenoj alternativnoj hipotezi $H_1 : \mu \neq \mu_0$.

To nam pokazuje da se općenito može smatrati da je problem nalaženja najboljeg testa, uza zadane n i α, riješen onda kada postoji jednoliko najsnažniji test. Nevolja je, međutim, u tome, što uvijek ne postoji jednoliko najsnažniji test za testiranje jednostavne hipoteze $H_0 : t = t_0$, prema složenoj alternativnoj hipotezi $H_1 : t \in \Theta_1$ ($\Theta_1 = \Theta \backslash \{t_0\}$), pa se prirodno nameće ideja da se usvoji novo načelo za definiranje najboljeg testa. To će nas dovesti do *metode omjera vjerodostojnosti*, odnosno do tzv. *LR-testova* (*likelihood ratio*).

5. Metoda omjera vjerodostojnosti

Već je ranije istaknuto opće načelo za definiranje kritičnog područja C nekog testa, koje se sastoji u tome da se u C uključe one točke $(x_1, \ldots, x_n) \in \mathbf{R}^n$ kojima, pod uvjetom da je hipoteza H_0 istinita, pripada mala vjerojatnost (gustoća vjerojatnosti) u usporedbi s vjerojatnosti te točke, uz uvjet da je istinita alternativna hipoteza H_1. To je načelo bilo jednostavno operacionalizirati u slučaju jednostavnih hipoteza, što je i učinjeno Neyman-Pearsonovom lemom, gdje je ključnu ulogu u definiranju najboljeg kritičnog područja imao omjer $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)}$. Mala vrijednost toga omjera u nekoj točki $(x_1, \ldots, x_n) \in \mathbf{R}^n$ upućivala je na veliku mogućnost da hipoteza H_0 nije istinita, što znači da tu točku treba uključiti u kritično područje testa.

Ako je riječ o testiranju jednostavne hipoteze $H_0 : t = t_0$, prema složenoj alternativnoj hipotezi $H_1 : t \in \Theta_1$, onda za svaki $t \in \Theta_1$ funkcija vjerodostojnosti ima određenu vrijednost $\boldsymbol{L}(t)$, tako da se ovdje ne može govoriti o omjeru $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t)}$ kao konstantnoj veličini pridruženoj točki $(x_1, \ldots, x_n) \in \mathbf{R}^n$. Taj omjer je sada funkcija nepoznatog parametra t. Postoji li međutim

$$(51)\quad \max_{t \in \Theta} \boldsymbol{L}(t) = \boldsymbol{L}(t_1),$$

onda je vrijednost omjera $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)}$ u točki $(x_1, \ldots, x_n)$ određeni fiksirani broj, pa ako je taj broj dovoljno malen, onda to upućuje da tu točku treba uključiti u kritično područje testa. Primijetimo najprije da je $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)} \leq 1$. Ako je $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)} \approx 0$, onda to označuje da je vjerojatnost (gustoća vjerojatnosti) $\boldsymbol{L}(t_0)$ da se dobije baš izmjereni niz podataka $x_1, \ldots, x_n$, uz uvjet da nepoznati parametar ima vrijednost t_0, zanemarivo mala u odnosu na najveću moguću vjerojatnost da se dobije taj niz podataka pri variranju parametra t po cijelom skupu Θ dopuštenih vrijednosti. Primijetimo, nadalje, da je $t_1 \in \Theta$ ona vrijednost parametra nepoznate vjerojatnosne razdiobe za koju dobiveni niz podataka ima najveću vjerojatnost (gustoću vjerojatnosti), pa je $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)}$ najmanja vrijednost omjera $\dfrac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t)}$ pri variranju parametra t po skupu Θ.

> Veličina
>
> $$(52)\quad \lambda(x_1, \ldots, x_n) = \frac{\boldsymbol{L}(t_0)}{\max\limits_{t \in \Theta} \boldsymbol{L}(t)} = \frac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)},$$
>
> zove se ***omjer vjerodostojnosti*** u točki $(x_1, \ldots, x_n) \in \mathbf{R}^n$.
>
> Ako je i nul-hipoteza složena hipoteza, tj. $H_0 : t \in \Theta_0$, i ako postoji
>
> $$(53)\quad \max_{t \in \Theta_0} \boldsymbol{L}(t) = \boldsymbol{L}(t_0),$$
>
> onda se omjer vjerodostojnosti definira formulom
>
> $$(54)\quad \lambda(x_1, \ldots, x_n) = \frac{\max\limits_{t \in \Theta_0} \boldsymbol{L}(t)}{\max\limits_{t \in \Theta} \boldsymbol{L}(t)} = \frac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)}.$$

Primijetimo da je t_1 vrijednost ML-procjenitelja za nepoznati parametar t (v. VI.3).

Očigledno je $\lambda(x_1, \ldots, x_n) \leq 1$, i ako se dobije $\lambda(x_1, \ldots, x_n) \approx 1$, onda točka $(x_1, \ldots, x_n)$ ne bi trebala pripadati kritičnom području za hipotezu H_0. Ako je pak $\lambda(x_1, \ldots, x_n) \approx 0$, onda niz mjerenja $x_1, \ldots, x_n$ upućuje na činjenicu da je njegova maksimalna vjerojatnost (gustoća vjerojatnosti) $\boldsymbol{L}(t_0)$, uz uvjet da je hipoteza H_0 istinita, zanemarivo mala u odnosu na njegovu maksimalno moguću vjerojatnost i da bi stoga točka $(x_1, \ldots, x_n)$ trebala pripadati kritičnom području hipoteze H_0.

Sada se čini razumnim smatrati da će se dobiti dobar test ako se kritično

područje C odabere tako da se u njega uključe one točke iz prostora $\mathbf{R}^n$ (prostor vrijednosti slučajnog uzorka) za koje je pripadni omjer vjerodostojnosti manji od zadanog broja c $(0 < c \leq 1)$.

Ako kritično područje C, za testiranje parametarske hipoteze $H_0 : t \in \Theta_0$, prema alternativnoj hipotezi $H_1 : t \in \Theta_1$, ima oblik

(55) $$C = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \lambda(x_1, \ldots, x_n) \leq c\},$$

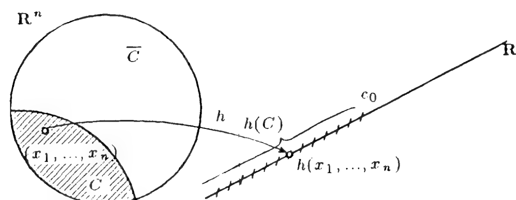
onda se kaže da je test dobiven ***metodom omjera vjerodostojnosti***, odnosno da je riječ o ***LR-testu***.

U formuli (55) $\lambda = \lambda(x_1, \ldots, x_n)$ ima značenje omjera vjerodostojnosti u točki $(x_1, \ldots, x_n)$. Ako se pretpostavi da nepoznati parametar vjerojatnosne razdiobe P_t ima konkretnu vrijednost $t \in \Theta$, onda se može govoriti o slučajnom vektoru $(X_1, \ldots, X_n)$ i statistici $\lambda(X_1, \ldots, X_n) = \Lambda$ kojoj pripada odgovarajuća funkcija razdiobe vjerojatnosti

(56) $$F(\lambda, t) = P_t(\Lambda \leq \lambda), \quad \lambda \in \mathbf{R}.$$

Često je statistika $\lambda(X_1, \ldots, X_n)$, koja proizlazi iz omjera vjerodostojnosti, takva da je teško odrediti njezinu razdiobu vjerojatnosti, pa se tada nejednakost $\lambda(x_1, \ldots, x_n) \leq c$, koja se pojavljuje u (55), transformira u ekvivalentnu nejednakost $h(x_1, \ldots, x_n) \leq c_0$ (ili $\geq c_0$), pri čemu statistika $h(X_1, \ldots, X_n) = Y_n$ obično ima lako odredivu razdiobu vjerojatnosti.

Statistika $Y_n = h(X_1, \ldots, X_n)$ zove se ***test-statistika*** danog testa. Test-statistikom se, zapravo, kritično područje $C \subseteq \mathbf{R}^n$ preslikava u skup realnih brojeva $\mathbf{R}$, kao odgovarajuće kritično područje $h(C) \subseteq \mathbf{R}$, što je geometrijski interpretirano na slici 20.



Slika 20. Interpretacija kritičnog područja na brojevnom pravcu

Želi li se, dakle, konstruirati LR-test, uz zadanu veličinu uzorka n i zadanu razinu značajnosti α, treba najprije pomoću omjera vjerodostojnosti naći funkciju h, a zatim odrediti vjerojatnosnu razdiobu statistike $Y_n = h(X_1, \ldots, X_n)$, uz pretpostavku da nepoznati parametar ima vrijednost t, tj. dobiti formulu za f.r.v. $F_n(y, t)$ slučajne varijable Y_n. To omogućuje da se odredi c_0 tako da kritično područje

(57) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : h(x_1, \ldots, x_n) \leq c_0\}$$

ima veličinu α, tj. da test ima razinu značajnosti α. Pripadna funkcija snage, prema (9), može se zapisati kao

(58) $$K(t) = P_t(h(X_1, \ldots, X_n) \leq c_0) = P_t(Y_n \leq c_0) = F_n(c_0, t), \quad t \in \Theta,$$

dok iz (12) proizlazi

(59) $$\alpha = \max_{t \in \Theta_0} F_n(c_0, t),$$

što načelno omogućuje nalaženje konstante c_0.

Lako se može provjeriti da u slučaju dvočlanog skupa $\Theta = \{t_0, t_1\}$ dopuštenih vrijednosti za parametar t, tj. kada je $H_0 : t = t_0$ i $H_1 : t = t_1$, metoda omjera vjerodostojnosti i Neyman-Pearsonova lema dovode do istog rezultata, iz čega se vidi da se načelo omjera vjerodostojnosti može smatrati određenom generalizacijom načela najmanje vjerojatnosti pogreške druge vrste.

Svi razmotreni primjeri, a to će se vidjeti i u idućim primjerima testova, pokazuju da vrijednost y_n test-statistike Y_n indicira stanje promatrane pojave u uvjetima istinitosti nul-hipoteze. Relacije (57) i (59) mogu se, prema tome, protumačiti tako da se nul-hipoteza odbacuje onda kada se dobije malo vjerojatna vrijednost y_n test-statistike Y_n, odnosno kada y_n padne u malo vjerojatno (vjerojatnosti ne veće od α) područje skupa svih mogućih vrijednosti test-statistike Y_n.

6. Testovi o parametrima normalne razdiobe

Metoda omjera vjerodostojnosti vrlo jednostavno omogućuje konstruiranje LR-testova za testiranje uobičajenih hipoteza o parametrima normalne razdiobe $N(\mu, \sigma^2)$. U idućim primjerima prikazat će se postupak određivanja kritičnog područja za određene hipoteze o parametrima normalne razdiobe uz pretpostavku da je zadana veličina uzorka n $(n \in \mathbf{N})$ i razina značajnosti α $(0 < \alpha < 1)$.

5. primjer

Pretpostavlja se da je σ $(\sigma > 0)$ poznato i treba konstruirati LR-test za testiranje jednostavne hipoteze $H_0 : \mu = \mu_0$, prema alternativnoj hipotezi $H_1 : \mu \neq \mu_0$ $(\mu \in \mathbf{R})$. Vodeći računa da je $t = \mu$ lako se uviđa da funkcija vjerodostojnosti glasi

(60) $$\boldsymbol{L}(\mu) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right], \quad \mu \in \mathbf{R}.$$

Lako se dokazuje (v. VI.4) da ona postiže svoj maksimum za $\mu = \overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$, tako da je

$$\max_{\mu \in \mathbf{R}} \boldsymbol{L}(\mu) = \boldsymbol{L}(\overline{x}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2\right].$$

U ovom je primjeru $t_0 = \mu_0$, pa se iz (52) vidi da omjer vjerodostojnosti glasi

$$\lambda(x_1,\ldots,x_n) = \frac{(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right]}{(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]} =$$
$$= \exp\left(\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2 - \sum_{i=1}^{n}(x_i-\mu_0)^2\right]\right) =$$
$$= \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)^2\right].$$

Kritično područje LR-testa definira se, prema (55), tako da se postavi uvjet

$$\exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)^2\right] \le c,$$

koji se, nakon logaritmiranja i sređivanja, može zapisati kao

(61) $$\left(\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right)^2 \ge -2\ln c.$$

Budući da je $0 < c \le 1$, onda je $\ln c \le 0$ i $\sqrt{-2\ln c} = c_0 \ge 0$, pa se uvjet (61) može pisati i kao

(62) $$\left|\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right| \ge c_0.$$

Stoga se kritično područje C_0 može zapisati u obliku

(63) $$C_0 = \left\{(x_1,\ldots,x_n)\in\mathbf{R}^n : \left|\frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}\right| \ge c_0\right\}.$$

Budući da je osnovna pretpostavka da mjerenja $x_1,\ldots,x_n$ potječu od slučajne varijable $X \sim \mathrm{N}(\mu,\sigma^2)$, onda statistika $\overline{X} \sim \mathrm{N}\left(\mu, \frac{1}{n}\sigma^2\right)$. Pretpostavljeno je, nadalje da je varijanca σ^2 poznata, pa ako je $\mu = \mu_0$, tj. kada je hipoteza H_0 stvarno istinita, statistika

(64) $$Z = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n} \sim \mathrm{N}(0,1).$$

Da bi razina značajnosti testa (veličina kritičnog područja C_0) iznosila α, na temelju (58), (59), (63) i (64) zaključuje se da treba biti

$$\alpha = P_0(|Z| \ge c_0) = 2 - 2\Phi(c_0),$$

odnosno

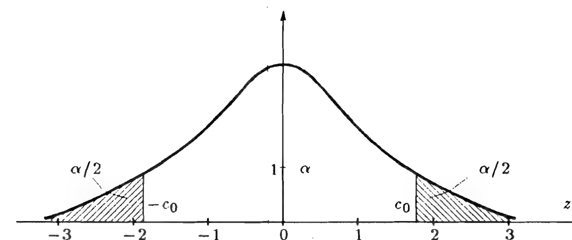
(65) $$c_0 = \Phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

Prema tome, LR-test za testiranje jednostavne hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \mu \ne \mu_0$, uz zadanu veličinu uzorka n i razinu značajnosti α, ima kritično područje

(66) $$C_0 = \{(x_1,\ldots,x_n)\in\mathbf{R}^n : |z| \ge c_0\},$$

gdje je $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$, a c_0 je određeno formulom (65). Vidi se da je C_0 određeno na temelju test-statistike Z, definirane u (64).

Uzme li se, na primjer, $\alpha = 0{,}05$, iz (65) se dobiva $c_0 = \Phi^{-1}(0{,}975) = 1{,}96$ (v. tabl. III. u Dodatku), što znači da se kritično područje može interpretirati kao skup točaka brojevnog pravca sastavljen od intervala $\langle-\infty; -1{,}96] \cup [1{,}96; \infty\rangle$, što je zorno prikazano na slici 21.



Slika 21. Skica kritičnog područja C_0 iz (66)

Dobije li se na uzorku veličine $n = 25$ vrijednost uzoračke aritmetičke sredine $\overline{x} = 4{,}6$, a testira li se hipoteza $\mathrm{H}_0 : \mu = 5$, prema alternativnoj hipotezi $\mathrm{H}_1 : \mu \ne 5$, pri čemu se pretpostavlja da normalna razdioba ima varijancu $\sigma^2 = 1$, dobit će se $|z| = \frac{|4{,}6-5|}{1}\cdot 5 = 2 > c_0 = 1{,}96$, pa se vidi da vrijednost test-statistike Z pada u kritično područje, što znači da hipotezu H_0 treba odbaciti. Dobivena vrijednost $|z| = 2$ pada u malo vjerojatno područje test-statistike Z, što upućuje na stanje u korist hipoteze H_1.

U modelu testiranja hipoteze o parametru μ normalne razdiobe $\mathrm{N}(\mu,\sigma^2)$, uz poznatu varijancu σ^2, mogu se razmatrati i druge različite hipoteze, kao što su, na primjer:

(a) $\mathrm{H}_0 : \mu = \mu_0$, prema $\mathrm{H}_1 : \mu > \mu_0$,

(b) $\mathrm{H}_0 : \mu = \mu_0$, prema $\mathrm{H}_1 : \mu < \mu_0$,

(c) $\mathrm{H}_0 : \mu \le \mu_0$, prema $\mathrm{H}_1 : \mu > \mu_0$,

(d) $\mathrm{H}_0 : \mu \ge \mu_0$, prema $\mathrm{H}_1 : \mu < \mu_0$.

Sličnim razmatranjem kao maloprije mogu se odrediti odgovarajuća kritična područja, uz zadano n i α, pri čemu se dobiva da u svakom od navedenih primjera (a)-(d) ključnu ulogu ima test-statistika Z, definirana u (64). Tako se pokazuje da odgovarajuća kritična područja pripadnih LR-testova glase:

(a_1) $$C_0 = \{(x_1,\ldots,x_n)\in\mathbf{R}^n : z \ge \Phi^{-1}(1-\alpha)\},$$

(b₁) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : z \le \Phi^{-1}(\alpha)\},$$

(c₁) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : z \ge \Phi^{-1}(1-\alpha)\},$$

(d₁) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : z \le \Phi^{-1}(\alpha)\},$$

gdje je $z = \dfrac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$ vrijednost test-statistike Z dobivena na n-članom slučajnom uzorku.

Imajući na umu razmatranja u 4. primjeru, vidi se da su testovi (a_1) i (b_1) također i jednoliko najsnažniji testovi.

U (64) se uočava da se vrijednosti test-statistike Z mogu interpretirati kao normirane razlike između istaknute vrijednosti μ_0 nepoznatog parametra μ i vrijednosti $\overline{x}$ procjenitelja $\overline{X}$ parametra μ. U slučaju (a) i (c) hipoteza H_0 se odbacuje kada se dobije prevelika normirana razlika (veća od $\Phi^{-1}(1-\alpha)$) između $\overline{x}$ i μ_0, tj. ako je dobiveno $\overline{x}$ previše desno od μ_0 na brojevnom pravcu. Tada, naime, izmjereni podaci upućuju na alternativnu hipotezu H_1 kao istinitu, što znači da H_0 treba odbaciti.

U slučaju (b) i (d) hipoteza H_0 se odbacuje kada se dobije premala (manja od $\Phi^{-1}(\alpha)$) normirana razlika između $\overline{x}$ i μ_0, tj. ako je dobiveno $\overline{x}$ previše lijevo od μ_0 na brojevnom pravcu.

Kao što smo već i ranije istaknuli, nul-hipoteza se odbacuje onda kada se dobije vrijednost test-statistike koja, u uvjetima istinitosti nul-hipoteze, pripada malo vjerojatnom području skupa svih mogućih vrijednosti test-statistike (v. sl. 21).

6. primjer

Parametar $\boldsymbol{t} = (\mu, \sigma^2)$ shvaća se kao vektorski parametar normalne razdiobe $N(\mu, \sigma^2)$, sa skupom dopuštenih vrijednosti $\boldsymbol{\Theta} = \{(\mu, \sigma^2) : \mu \in \mathbf{R}, \sigma^2 > 0\}$. Istakne li se njegov podskup $\boldsymbol{\Theta}_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}$, gdje je μ_0 fiksirani realni broj, može se postaviti zadatak testiranja složene hipoteze $H_0 : \boldsymbol{t} \in \boldsymbol{\Theta}_0$, prema alternativnoj složenoj hipotezi $H_1 : \boldsymbol{t} \in \boldsymbol{\Theta}_1$ ($\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} \backslash \boldsymbol{\Theta}_0$). Uobičajeno je da se taj zadatak formulira kao testiranje hipoteze $H_0 : \mu = \mu_0$, prema alternativnoj hipotezi $H_1 : \mu \ne \mu_0$, uz pretpostavku da je varijanca σ^2 nepoznata.

Funkcija vjerodostojnosti zapisana je u formuli (45) u VI.4, gdje je riješen i problem određivanja $\max_{\boldsymbol{t}\in\boldsymbol{\Theta}} \boldsymbol{L}(\boldsymbol{t}) = \boldsymbol{L}(\boldsymbol{t}_1)$, pri čemu je $\boldsymbol{t}_1 = (\overline{x}, \widehat{\sigma}^2)$, $\left(\widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)$.

Iz definicije skupa $\boldsymbol{\Theta}_0$ proizlazi da je

$$\max_{\boldsymbol{t}\in\boldsymbol{\Theta}_0} \boldsymbol{L}(\boldsymbol{t}) = \max_{\sigma^2>0} \boldsymbol{L}(\mu_0, \sigma^2) = \boldsymbol{L}(\mu_0, \sigma_0^2),$$

gdje je $\sigma_0^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$, tako da pripadni omjer vjerodostojnosti glasi

$$\lambda(x_1, \ldots, x_n) = \frac{\boldsymbol{L}(\boldsymbol{t}_0)}{\boldsymbol{L}(\boldsymbol{t}_1)} = \left(\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^{-\frac{n}{2}}.$$

Kritično područje pripadnog LR-testa određeno je uvjetom

$$\left(\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^{-\frac{n}{2}} \le c,$$

koji nakon sređivanja postaje

(67) $$\frac{(\overline{x} - \mu_0)^2}{\widehat{\sigma}^2} \ge c^{-\frac{2}{n}} - 1.$$

Uzme li se još u obzir da je $\widehat{\sigma}^2 = \dfrac{n-1}{n}s^2$ ($s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$), (67) postaje

$$\frac{n}{n-1}\frac{(\overline{x} - \mu_0)^2}{s^2} \ge c^{-\frac{n}{2}} - 1,$$

odnosno

(68) $$\left|\frac{\overline{x} - \mu_0}{s}\sqrt{n}\right| \ge c_0, \quad c_0 = \sqrt{(n-1)(c^{-\frac{2}{n}} - 1)} \ge 0.$$

Kritično područje C_0 LR-testa za testiranje hipoteze $H_0 : \mu = \mu_0$, prema alternativnoj hipotezi $H_1 : \mu \ne \mu_0$, pri nepoznatoj varijanci σ^2 normalne razdiobe $N(\mu, \sigma^2)$, može se, prema tome, zapisati u obliku

(69) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \left|\frac{\overline{x} - \mu_0}{s}\sqrt{n}\right| \ge c_0\}.$$

Ostaje još da se konstanta c_0 odredi tako da kritično područje C_0 ima zadanu veličinu α. U tu svrhu primijetimo najprije da se $\overline{x}$ i s^2 mogu shvatiti kao vrijednosti statistika $\overline{X}$ (uzoračka aritmetička sredina) i S^2 (uzoračka korigirana varijanca), pa se može govoriti o test-statistici $Y_n = T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$. Uočimo također da, u uvjetima istinitosti hipoteze H_0, statistici T pripada Studentova razdioba $t(n-1)$ sa $n-1$ stupnjeva slobode (v. VI.2). Može se, stoga, pisati

(70) $$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n} \sim t(n-1).$$

Budući da je

$$\alpha = P_0((X_1, \ldots, X_n) \in C_0) = P_0(|T| \geq c_0),$$

pa ako se još sa G_n označi f.r.v. Studentove razdiobe sa n stupnjeva slobode, može se pisati

$$P_0(|T| \geq c_0) = 2 - 2G_{n-1}(c_0),$$

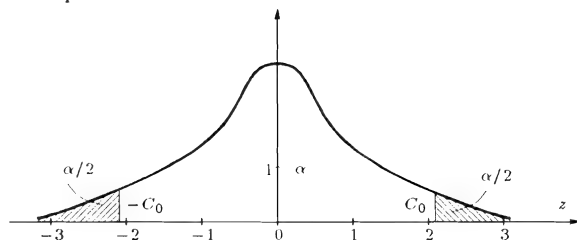
što proizlazi iz činjenice da je Studentova razdioba simetrična s obzirom na ishodište (v. sl. 22). Konačno se dobiva

$$c_0 = G_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right), \tag{71}$$

gdje je G_n^{-1} inverzna funkcija od G_n.

Iz (69) i (71) razabire se da kritično područje C_0 ovisi, kao i uvijek, o veličini uzorka n i razini značajnosti α, te da se odluka o prihvaćanju, odnosno odbacivanju, hipoteze H_0 donosi na temelju vrijednosti $\tau = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$ test-statistike T, opisane u (70). Zanimljivo je primijetiti da se i u ovom slučaju vrijednost test-statistike može interpretirati kao određena normirana razlika između istaknute vrijednosti μ_0 parametra μ i vrijednosti $\overline{x}$ njegova procjenitelja $\overline{X}$.

Uzme li se, na primjer, $\alpha = 0{,}05$ i $n = 25$, primjenom tabl. V. iz Dodatka, dobiva se $c_0 = G_{24}^{-1}\left(1 - \dfrac{0{,}05}{2}\right) = G_{24}^{-1}(0{,}975) = 2{,}064$, pa se kritično područje može prikazati kao skup točaka brojevnog pravca sastavljen od intervala $\langle -\infty\,; -2{,}064] \cup [2{,}064\,; \infty\rangle$, što je prikazano na slici 22, zajedno s krivuljom razdiobe test-statistike T. Dobiju li se na izmjerenim podacima vrijednosti $\overline{x} = 4{,}6$ i $s = 1$, a testira se hipoteza $H_0 : \mu = 5$, prema alternativnoj hipotezi $H_1 : \mu \neq 5$, vrijednost je test-statistike $\tau = \dfrac{4{,}6 - 5}{1} \cdot 5 = -2$. Budući da je $|\tau| = 2 < c_0 = 2{,}064$, hipotezu H_0 treba prihvatiti.



Slika 22. Skica kritičnog područja C_0 iz (69)

Usporedi li se ovaj rezultat s onim u 5. primjeru (usp. sl. 21. i 22), razabire se da je kritično područje u 5. primjeru, uz iste n i α, nešto opsežnije od ovoga u 6. primjeru. No, to se moglo i očekivati, jer se u 5. primjeru pretpostavlja poznavanje varijance σ^2, dok se u 6. primjeru primjenjuje procjena s^2 za nepoznatu varijancu σ^2, što utječe na veći oprez pri odbacivanju hipoteze H_0, tj. na smanjenje kritičnog područja. Zato se i moglo dogoditi da se, na temelju jednakih vrijednosti

odgovarajućih test-statistika ($z = \tau = 2$), jednom hipoteza H_0 odbaci (5. primjer), a drugi put prihvati (6. primjer).

Istaknimo još i to da se hipoteze (a)-(d) iz 5. primjera mogu testirati i uz pretpostavku da je varijanca σ^2 nepoznata. Može se dokazati (v. zad. 12) da se pripadna kritična područja odgovarajućih LR-testova također mogu definirati pomoću vrijednosti τ statistike T, definirane u (70). Ona glase:

(a₂) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \tau \geq G_{n-1}^{-1}(1-\alpha)\},$$

(b₂) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \tau \leq G_{n-1}^{-1}(\alpha)\},$$

(c₂) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \tau \geq G_{n-1}^{-1}(1-\alpha)\},$$

(d₂) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \tau \leq G_{n-1}^{-1}(\alpha)\},$$

gdje je $\tau = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$ vrijednost test-statistike T.

7. primjer

Pretpostavlja se da je $\mu = 0$ i treba, primjenom Neyman-Pearsonove leme, konstruirati test za testiranje jednostavne hipoteze $H_0 : \sigma^2 = \sigma_0^2$, prema alternativnoj hipotezi $H_1 : \sigma^2 = \sigma_1^2$, gdje su σ_0 i σ_1 zadani pozitivni brojevi.

Stavimo $t = \sigma^2$, pa je očigledno da pripadna funkcija vjerodostojnosti glasi

$$\boldsymbol{L}(t) = \boldsymbol{L}(\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right). \tag{72}$$

Nadalje je

$$\frac{\boldsymbol{L}(t_0)}{\boldsymbol{L}(t_1)} = \frac{\boldsymbol{L}(\sigma_0^2)}{\boldsymbol{L}(\sigma_1^2)} = \left(\frac{\sigma_0}{\sigma_1}\right)^{-n} \exp\left(-\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\sum_{i=1}^{n} x_i^2\right),$$

što pokazuje da se najbolje kritično područje C_0 dobiva iz uvjeta

$$\left(\frac{\sigma_1}{\sigma_0}\right)^{n} \exp\left(-\frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2}\sum_{i=1}^{n} x_i^2\right) \leq c. \tag{73}$$

Pretpostavimo da je $\sigma_1 > \sigma_0$, pa se logaritmiranjem relacije (73) dobiva

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i^2 \geq \frac{2\sigma_1^2}{\sigma_1^2 - \sigma_0^2}\left(n \ln\frac{\sigma_1}{\sigma_0} - \ln c\right) = c_0,$$

iz čega proizlazi da se kritično područje može zapisati u obliku

$$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i^2 \geq c_0\}. \tag{74}$$

Budući da je polazna pretpostavka da mjerenja $x_1, \ldots, x_n$ potječu od slučajne varijable $X \sim \mathrm{N}(0, \sigma^2)$, test-statistika $U = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2$ ima hikvadrat-razdiobu sa n stupnjeva slobode (v. točku 5. u V.6). Da bi razina značajnosti testa s kritičnim područjem (74) iznosila α mora vrijediti

$$\alpha = P_0((X_1, \ldots, X_n) \in C_0) = P_0(U \geq c_0). \tag{75}$$

Označi li se sa H_n f.r.v. hikvadrat-razdiobe $\chi^2(n)$, može se pisati

$$P_0(U \geq c_0) = 1 - P_0(U < c_0) = 1 - H_n(c_0),$$

pa (75) postaje

$$\alpha = 1 - H_n(c_0),$$

iz čega se dobiva

$$c_0 = H_n^{-1}(1 - \alpha), \tag{76}$$

gdje je H_n^{-1} inverzna funkcija od H_n.

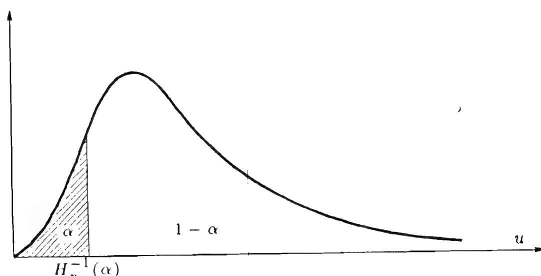
Iz (74) i (76) vidi se da najbolji test za testiranje jednostavne hipoteze $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$, prema jednostavnoj alternativnoj hipotezi $\mathrm{H}_1 : \sigma^2 = \sigma_1^2$ $(\sigma_1 > \sigma_0)$, uz zadano n i α, ima kritično područje

$$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : u \geq c_0\}, \tag{77}$$

gdje je $u = \frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_i^2$ vrijednost test-statistike $U \sim \chi^2(n)$, a c_0 je određeno formulom (76).

Budući da kritično područje C_0 ne ovisi o pretpostavljenoj vrijednosti σ_1^2, znači da je dobiveni test jednoliko najsnažniji test za testiranje hipoteze $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \sigma^2 > \sigma_0^2$.

Tako se, na primjer, za $n = 25$ i $\alpha = 0{,}05$, primjenom tabl. VI. iz Dodatka, dobiva da je $c_0 = H_{25}^{-1}(0{,}95) = 37{,}7$, pa se pripadno kritično područje može interpretirati kao interval $[37{,}7; \infty\rangle$ brojevnog pravca. Testira li se hipoteza $\mathrm{H}_0 : \sigma^2 = 10$, prema alternativnoj hipotezi $\mathrm{H}_1 : \sigma^2 > 10$, i na uzorku od 25 mjerenja dobije zbroj kvadrata $\sum_{i=1}^{25} x_i^2 = 260$, vrijednost test-statistike U bit će $u = 26$. Kako je $26 < 37{,}7$,



Slika 23. Skica kritičnog područja C_0 iz (78)

vrijednost test-statistike ne pada u kritično područje, a to znači da hipotezu H_0 treba prihvatiti. Smatra se da dobiveni zbroj kvadrata izmjerenih podataka nije prevelik u usporedbi s hipotetičkom varijancom $(\sigma_0^2 = 10)$ pretpostavljene normalne razdiobe, što upućuje na prihvaćanje hipoteze H_0.

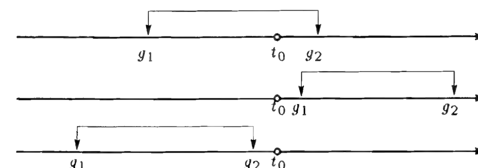
Sličnim razmatranjima može se izvesti da je

$$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : u \leq H_n^{-1}(\alpha)\} \tag{78}$$

kritično područje jednoliko najsnažnijeg testa za testiranje jednostavne hipoteze $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \sigma^2 < \sigma_0^2$ (v. zad. 10b).

7. Primjena intervala povjerenja

Konstruiranje najboljih testova primjenom Neyman-Pearsonove leme i metode omjera vjerodostojnosti često je računski vrlo složeno tako da se u mnogim, za praksu važnim, situacijama ne može dobiti dovoljno jednostavno rješenje. Imajući na umu pojam i svojstva intervala povjerenja zadane pouzdanosti γ, prirodno se nameće ideja da se pri konstrukciji kritičnog područja C_0, razine značajnosti $\alpha = 1 - \gamma$, za testiranje jednostavne hipoteze $\mathrm{H}_0 : t = t_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : t \neq t_1$, postupi na sljedeći način: Na temelju niza mjerenja $x_1, \ldots, x_n$ odredi se interval povjerenja $\langle g_1, g_2 \rangle$, pouzdanosti γ, za nepoznati parametar t. Pokrije li taj interval točku t_0 brojevnog pravca, tj. ako je $g_1 < t_0 < g_2$, hipoteza H_0 se prihvaća, a u protivnom se odbacuje.



Slika 24. Skica primjene intervala povjerenja u testiranju hipoteza

To znači da je pripadno kritično područje C_0 oblika

$$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : g_1 \geq t_0 \text{ ili } g_2 \leq t_0\}. \tag{79}$$

Budući da su, u uvjetima istinitosti hipoteze H_0, g_1 i g_2 vrijednosti statistika G_1 i G_2 (v. VII.1), za koje vrijedi $P_0(G_1 < t_0 < G_2) \geq \gamma$, i imajući na umu definiciju razine značajnosti testa (v. (9) i (14)), očigledno je da kritičnom području (79) pripada razina značajnosti $\alpha \leq 1 - \gamma$. Može se, naime, pisati

$$\alpha = P_0((X_1, \ldots, X_n) \in C_0) = 1 - P_0(G_1 < t_0 < G_2) \leq 1 - \gamma. \tag{80}$$

Ako statistike G_1 i G_2 određuju najuži interval povjerenja pouzdanosti γ za nepoznati parametar t, onda u (80) vrijedi znak jednakosti.

Prema tome, opisanim se postupkom dobiva test sa zadanom veličinom uzorka n i razinom značajnosti α, ali se ne dobiva odgovor na pitanje o "kvaliteti" toga

testa. U nekim slučajevima, posebno kada je riječ o normalnoj razdiobi, opisana metoda primjene intervala povjerenja i metoda omjera vjerodostojnosti dovode do istoga kritičnog područja (v. zad. 13).

8. primjer

Uz pretpostavku da mjerenja $x_1, \dots, x_n$ potječu od normalne razdiobe $N(\mu, \sigma^2)$, treba konstruirati test sa zadanim n i α za testiranje hipoteze $H_0 : \sigma^2 = \sigma_0^2$, prema alternativnoj hipotezi $H_1 : \sigma^2 \neq \sigma_0^2$.

Primijetimo najprije da je u VII.2. riješen problem određivanja najužeg intervala povjerenja pouzdanosti γ za nepoznatu varijancu σ^2 normalne razdiobe $N(\mu, \sigma^2)$, čije su granice određene formulama (30) u VII.2. Iz toga i (79) slijedi da će traženo kritično područje biti određeno nejednakošću

$$\frac{n-1}{u_2}s^2 \geq \sigma_0^2 \implies (n-1)\frac{s^2}{\sigma_0^2} \geq u_2,$$

odnosno nejednakošću

$$\frac{n-1}{u_1}s^2 \leq \sigma_0^2 \implies (n-1)\frac{s^2}{\sigma_0^2} \leq u_1,$$

gdje je s^2 vrijednost korigirne uzoračke varijance, a u_1 i u_2 odgovarajuće vrijednosti ovisne o zadanoj pouzdanosti γ, odnosno razini značajnosti α. Vidi se da odlučujuću ulogu u definiranju kritičnog područja ima vrijednost $u = (n-1)\frac{s^2}{\sigma_0^2}$ statistike

(81) $$U = (n-1)\frac{S^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Prema tome, kritično područje C_0 pri testiranju hipoteze $H_0 : \sigma^2 = \sigma_0^2$, prema alternativnoj hipotezi $H_1 : \sigma^2 \neq \sigma_0^2$, glasi

(82) $$C_0 = \{(x_1, \dots, x_n) \in \mathbf{R}^n : u \notin \langle u_1, u_2 \rangle\},$$

gdje je

(83) $$u_1 = H_{n-1}^{-1}\left(\frac{\alpha}{2}\right), \quad u_2 = H_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Ideje metode intervala povjerenja mogu se primjeniti i na određivanje kritičnog područja u tzv. *jednorubnim testovima*, tj. takvima gdje se testira jednostavna hipoteza $H_0 : t = t_0$, prema alternativnoj hipotezi $H_1 : t < t_0$ (ili $t > t_0$). Može se, naime, poći od toga da se za nepoznati parametar t, dopuštene klase vjerojatnosnih razdioba $\mathcal{P} = \{P_t : t \in \Theta\}$, najprije nađe tzv. *jednorubni interval povjerenja* zadane pouzdanosti γ $(0 < \gamma < 1)$, tj. takva statistika $\widetilde{G}_2$ za koju vrijedi

(84) $$P_t(\widetilde{G}_2 > t) = \gamma,$$

odnosno takva statistika $\widetilde{G}_1$ za koju vrijedi

(85) $$P_t(\widetilde{G}_1 < t) = \gamma.$$

Relacija (84) može se interpretirati tako da se kaže da interval $\langle -\infty, \widetilde{G}_2 \rangle$ pokriva nepoznati parametar t s vjerojatnošću γ. Isto se tako (85) može protumačiti da interval $\langle \widetilde{G}_1, \infty \rangle$ pokriva nepoznati parametar t s vjerojatnošću γ. U uvjetima istinitosti hipoteze H_0, (84) i (85) postaju

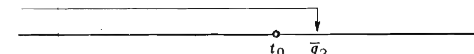
(84a), $$P_0(\widetilde{G}_2 > t_0) = \gamma,$$

odnosno

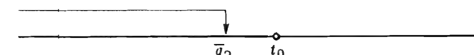
(85a), $$P_0(\widetilde{G}_1 < t_0) = \gamma.$$

Kada je alternativna hipoteza $H_1 : t < t_0$, kritično područje C_0 definirat će se nejednakošću $\tilde{g}_2 < t_0$, a kada je $H_1 : t > t_0$, nejednakošću $\tilde{g}_1 > t_0$, gdje su $\tilde{g}_1$ i $\tilde{g}_2$ vrijednosti statistika $\widetilde{G}_1$ i $\widetilde{G}_2$ na n-članom slučajnom uzorku. Dobije li se, dakle, jednorubni interval povjerenja pouzdanosti $\gamma = 1 - \alpha$, koji pokriva točku t_0 brojevnog pravca, hipoteza H_0 se prihvaća, dok se u protivnom prihvaća alternativna hipoteza H_1.

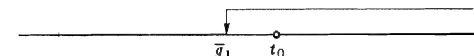
a) prihvaća se $H_0 : t = t_0$
odbacuje se $H_1 : t < t_0$



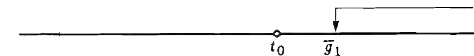
b) odbacuje se $H_0 : t = t_0$
prihvaća $H_1 : t < t_0$



c) prihvaća se $H_0 : t = t_0$
odbacuje se $H_1 : t > t_0$



d) odbacuje se $H_0 : t = t_0$
prihvaća se $H_1 : t > t_0$



Slika 25. Skica primjene jednorubnih intervala povjerenja u testiranju jednorubnih hipoteza

Budući da pri $H_1 : t < t_0$ imamo

$$\alpha = P_0((X_1, \dots, X_n) \in C_0) = P_0(\widetilde{G}_2 < t_0) = 1 - \gamma,$$

dok pri $H_1 : t > t_0$ imamo

$$\alpha = P_0((X_1, \dots, X_n) \in C_0) = P_0(\widetilde{G}_1 > t_0) = 1 - \gamma,$$

vidi se da tako konstruirani test ima razinu značajnosti $\alpha = 1 - \gamma$.

Može se konačno reći da je

(86) $$C_0 = \{(x_1, \dots, x_n) \in \mathbf{R}^n : \tilde{g}_2 \leq t_0 \ (\tilde{g}_1 \geq t_0)\}$$

kritično područje razine značajnosti α za testiranje jednostavne hipoteze $H_0 : t = t_0$, prema alternativnoj hipotezi $H_1 : t > t_0$ $(t < t_0)$, pri čemu je $\tilde{g}_i$ vrijednost test-statistike $\widetilde{G}_i$ $(i = 1, 2)$.

Preformulira li se 8. primjer tako da se za alternativnu hipotezu uzme $H_1 : \sigma^2 < \sigma_0^2$ (ili $\sigma^2 > \sigma_0^2$), na temelju onog što je navedeno u VII.2, vidi se da je

$$\tilde{G}_2 = \frac{n-1}{\tilde{u}_1}S^2 \quad \left(\tilde{G}_1 = \frac{n-1}{\tilde{u}_2}S^2\right),$$

gdje je S^2 korigirana uzoračka varijanca i

(87) $$\tilde{u}_1 = H_{n-1}^{-1}(\alpha), \quad \tilde{u}_2 = H_{n-1}^{-1}(1-\alpha).$$

To znači da je odgovarajuće kritično područje C_0 određeno uvjetom

$$\tilde{g}_2 = \frac{n-1}{\tilde{u}_1}s^2 \le \sigma_0^2 \quad \left(\tilde{g}_1 = \frac{n-1}{\tilde{u}_2}s^2 \ge \sigma_0^2\right),$$

što se može pisati i kao

$$\frac{n-1}{\sigma_0^2}s^2 \le \tilde{u}_1 \quad \left(\frac{n-1}{\sigma_0^2}s^2 \ge \tilde{u}_2\right).$$

Može se, konačno, reći da je

(88) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : u \le \tilde{u}_1 \ (u \ge \tilde{u}_2)\}$$

kritično područje razine značajnosti α jednorubnog testa $H_0 : \sigma^2 = \sigma_0^2$, prema $H_1 : \sigma^2 < \sigma_0^2$ ($\sigma^2 > \sigma_0^2$), pri čemu je $u = \frac{n-1}{\sigma_0^2}s^2$ vrijednost test-statistike

(89) $$U = \frac{n-1}{\sigma_0^2}S^2 \sim \chi^2(n-1).$$

Činjenica koja je već istaknuta u VIII.3, da se za velike n interval povjerenja nepoznatog parametra t može odrediti primjenom asimptotske normalnosti odgovarajućeg procjenitelja omogućuje da se konstruiraju određeni testovi i bez konkretiziranja tipa vjerojatnosne razdiobe.

Neka je, dakle, n veliko i $\hat{T}_n$ nepristran, konzistentan i asimptotski normalan procjenitelj za parametar t, tada približno vrijedi $\hat{T}_n \sim N(t, R_n(t))$, gdje je $R_n(t) = V[\hat{T}_n]$. Treba li konstruirati test za testiranje jednostavne hipoteze $H_0 : t = t_0$, prema alternativnoj hipotezi $H_1 : t \ne t_0$, uzet će se

(90) $$Z_n = \frac{\hat{T}_n - t_0}{\sqrt{R_n(t_0)}}$$

kao test-statistika i kritično područje C_0 odredit će se uvjetom $|z_n| > c_0$, pri čemu će se konstanta c_0 odrediti tako da vrijedi

(91) $$P_0(|Z_n| > c_0) = \alpha,$$

tj. tako da test ima razinu značajnosti α.

Budući da približno vrijedi $Z_n \sim N(0,1)$, iz (91) proizlazi

$$c_0 = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right),$$

tako da kritično područje C_0 glasi

(92) $$C_0 = \left\{(x_1, \ldots, x_n) \in \mathbf{R}^n : |z_n| \ge \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right\},$$

gdje je

(93) $$z_n = \frac{\hat{t}_n - t_0}{\sqrt{R_n(t_0)}}.$$

Uzme li se $H_1 : t < t_0$ ($t > t_0$) kao alternativna hipoteza, kritično područje C_0 odredit će se uvjetom $z_n < c_1$ ($z_n > c_2$), pri čemu će se konstanta $c_1(c_2)$ odrediti tako da test ima razinu značajnosti α, tj. da vrijedi

(94) $$P_0(Z_n < c_1) = \alpha \quad (P_0(Z_n > c_2) = \alpha).$$

Iz (94) i činjenice da $Z_n \sim N(0,1)$ proizlazi

$$c_1 = \Phi^{-1}(\alpha) \quad (c_2 = \Phi^{-1}(1-\alpha)),$$

tako da pripadno kritično područje glasi

(95) $$C_0 = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : z_n \le \Phi^{-1}(\alpha) \quad (z_n \ge \Phi^{-1}(1-\alpha))\}.$$

Uzme li se, na primjer, očekivanje μ kao nepoznati parametar ($t = \mu$), uz pretpostavku da je poznata varijanca σ^2, tada će se iskoristiti činjenica da je $\hat{T}_n = \overline{X}$ nepristran, konzistentan i asimptotski normalan procjenitelj za nepoznato očekivanje μ i da je $R_n(t) = V[\overline{X}] = \frac{1}{n}\sigma^2$, tako da je

(96) $$Z_n = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}$$

odgovarajuća test-statistika i sa (95), odnosno (92), određeno je kritično područje razine značajnosti α za testiranje hipoteze $H_0 : \mu = \mu_0$, prema odgovarajućoj alternativnoj hipotezi H_1.

Usporedi li se to s rezultatima iz 5. primjera (v. (64), (65) i (66), te (a_1) i (b_1)), razabire se da je riječ o istim formulama. Razlika je u tome što formule u 5. primjeru vrijede za svaki $n > 1$, uz pretpostavku da mjerenja potječu od normalne razdiobe, dok (92) i (95) vrijede za velike n, uz pretpostavku da mjerenja potječu od vjerojatnosne razdiobe s konačnom varijancom σ^2.

U prethodnim razmatranjima pretpostavljeno je da je t jednodimenzionalni parametar, međutim slična se procedura pri konstrukciji testa može primijeniti i kada je riječ o vektorskom parametru $\boldsymbol{t}$. Uzmimo, na primjer, da je $\boldsymbol{t} = (\mu, \sigma^2)$ dvodimenzionalni vektorski parametar s komponentama $t_1 = \mu$ (očekivanje) i $t_2 = \sigma^2$ (varijanca) i da treba konstruirati test razine značajnosti α za testiranje

hipoteze $\mathrm{H}_0 : \mu = \mu_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \mu \neq \mu_0$ (ili $\mu < \mu_0$, ili $\mu > \mu_0$), u uvjetima nepoznate varijance σ^2. Očigledno je da se sada ne može uzeti Z_n iz (96) kao test-statistika, jer se u njoj pojavljuje napoznata druga komponenta $t_2 = \sigma^2$. Stoga će se σ^2 zamijeniti nepristranim i konzistentnim procjeniteljem S^2, čime se dobiva statistika

$$\widehat{Z}_n = \frac{\overline{X} - \mu_0}{S}\sqrt{n}, \tag{97}$$

za koju je već rečeno (v. VII.3) da za velike n približno ima standardnu normalnu razdiobu $\mathrm{N}(0,1)$.

Prema tome, i u ovom je slučaju kritično područje određeno formulama (92), odnosno (95), samo što umjesto z_n treba staviti vrijednost $\widehat{z}_n = \dfrac{\overline{x} - \mu_0}{s}\sqrt{n}$ test-statistike $\widehat{Z}_n$ iz (97).

Promotre li se hipoteze o drugoj komponenti $t_2 = \sigma^2$, tako da se uzme $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$, uz alternativnu hipotezu $\mathrm{H}_1 : \sigma^2 \neq \sigma_0^2$ (ili $\sigma^2 < \sigma_0^2$, ili $\sigma^2 > \sigma_0^2$), iskoristit će se činjenica da je korigirana uzoračka varijanca S^2 nepristran i asimptotski normalan procjenitelj za nepoznatu varijancu σ^2. Budući da je $\mathrm{E}[S^2] = \sigma^2$ i $\mathrm{V}[S^2] = \dfrac{1}{n}\left(\varepsilon + \dfrac{2n}{n-1}\right)\sigma^4$ (v. VI. 2. i 5. primjer u VII.3), za velike n, u uvjetima istinitosti hipoteze H_0, približno vrijedi

$$\begin{aligned} \widetilde{Z}_n = \frac{S^2 - \mathrm{E}[S^2]}{\mathrm{V}[S^2]} &= \left(\frac{S^2}{\sigma_0^2} - 1\right)\sqrt{\frac{n(n-1)}{(n-1)\varepsilon + 2n}} \approx \\ &\approx \left(\frac{S^2}{\sigma_0^2} - 1\right)\sqrt{\frac{n}{\varepsilon + 2}} \sim \mathrm{N}(0,1). \end{aligned} \tag{98}$$

To znači da će se vrijednost $\widetilde{z}_n = \left(\dfrac{s^2}{\sigma_0^2} - 1\right)\sqrt{\dfrac{n}{\varepsilon + 2}}$ test-statistike $\widetilde{Z}_n$ moći konkretno izračunati onda kada je poznata vrijednost koeficijenta spljoštenosti ε i vrijednost s^2 korigirane uzoračke varijance.

Ako se može pretpostaviti da je $\varepsilon = 0$, tj. da je spljoštenost pretpostavljene klase vjerojatnosnih razdioba ista kao normalne razdiobe, onda (98) postaje

$$\widetilde{Z}_n = \left(\frac{S^2}{\sigma_0^2} - 1\right)\sqrt{\frac{n-1}{2}}. \tag{99}$$

Ako ε nije poznato, može se u (98), umjesto ε, staviti neki procjenitelj $\widehat{\varepsilon}$ za nepoznati parametar ε (uzme se, recimo, uzorački koeficijent spljoštenosti kao $\widehat{\varepsilon}$).

Zanimljivo je primijetiti da vrijednost $\widetilde{z}_n$ test-statistike $\widetilde{Z}_n$ iz (98) i (99) upućuje na veličinu razlike između jedinice i omjera mjere rasipanja uzoračkih podataka (s^2) i mjere rasipanja pretpostavljene vjerojatnosne razdiobe (σ_0^2). Prevelike vrijednosti te razlike upućuju na odbacivanje hipoteze $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$.

9. primjer

Prethodno zapažanje nameće nam ideju da se sličnim zaključivanjem konstruira test, zadane razine značajnosti α, za testiranje hipoteze $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$, prema alternativnoj hipotezi $\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$ (ili $\sigma_1^2 < \sigma_2^2$), pri čemu se pretpostavlja da imamo niz mjerenja $x_1, \dots, x_m$, koja potječu od slučajne varijable $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, i niz mjerenja $y_1, \dots, y_n$, koja potječu od slučajne varijable $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, te da su X i Y nezavisne slučajne varijable.

Takva hipoteza najčešće se u praksi pojavljuje pri kontroli ujednačenosti (stabilnosti) određenih tehnoloških procesa. Proces se obično kontrolira u određenim vremenskim razmacima, tako da se načini određeni broj mjerenja relevantne veličine. Ako dva takva niza mjerenja dopuštaju zaključak da potječu od teorijskih razdioba vjerojatnosti iste varijance, smatra se da je proces stabilan. U protivnom se smatra da su nastale značajne promjene u odvijanju procesa, što obično zahtijeva određenu tehnološku intervenciju.

Polazeći od činjenice da su korigirane uzoračke varijance S_x^2 i S_y^2 dobri procjenitelji za nepoznate parametre σ_1^2 i σ_2^2, odmah se nameće ideja da bi statistika $\dfrac{S_x^2}{S_y^2}$ mogla poslužiti kao osnova za konstrukciju traženog testa. Poznato je, naime, da

$$U_1 = \frac{m-1}{\sigma_1^2}S_x^2 \sim \chi^2(m-1), \quad U_2 = \frac{n-1}{\sigma_2^2}S_y^2 \sim \chi^2(n-1),$$

te da su U_1 i U_2 nezavisne slučajne varijable, pa se prema točki 8. u V.6. zaključuje da slučajna varijabla

$$V = \frac{(n-1)U_1}{(m-1)U_2} = \frac{\sigma_2^2}{\sigma_1^2} \cdot \frac{S_x^2}{S_y^2}$$

ima F-razdiobu sa $(m-1, n-1)$ stupnjeva slobode.

U uvjetima istinitosti hipoteze $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$ govorit će se o test-statistici

$$V = \frac{S_x^2}{S_y^2} \sim \mathrm{F}(m-1, n-1). \tag{100}$$

Vrijednost test-statistike V na danim mjerenjima je

$$v = \frac{s_x^2}{s_y^2} = \frac{n-1}{m-1} \cdot \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \geq 0, \tag{101}$$

pa mala vrijednost test-statistike (bliska nuli) indicira da bi σ_1^2 moglo biti manje od σ_2^2, dok velika vrijednost (mnogo veća od jedan) indicira da bi moglo biti $\sigma_1^2 > \sigma_2^2$. Ako je, pak, v blisko jedinici, onda to upućuje na zaključak da je $\sigma_1^2 = \sigma_2^2$. Stoga će se kritično područje za testiranje hipoteze $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$, prema alternativnoj hipotezi $\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$, definirati uvjetima

$$v \leq c_1, \quad (0 < c_1 < 1), \qquad v \geq c_2, \quad (1 < c_2 < \infty),$$

pri čemu će se c_1 i c_2 odrediti iz zahtjeva da test ima razinu značajnosti α, tj. iz jednadžbi

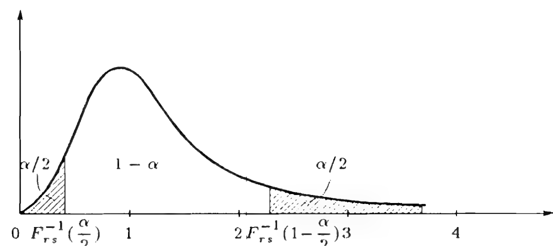
$$P_0(V < c_1) = P_0(V > c_2) = \frac{\alpha}{2}. \tag{102}$$

Označi li se sa F_{rs} f.r.v. F-razdiobe sa (r,s) stupnjeva slobode, F_{rs}^{-1} će označavati odgovarajuću inverznu funkciju, pa iz (102) odmah slijedi

$$c_1 = F_{rs}^{-1}\left(\frac{\alpha}{2}\right), \quad c_2 = F_{rs}^{-1}\left(1-\frac{\alpha}{2}\right), \quad r = m-1, \quad s = n-1.$$

Prema tome, kritično područje testa, razine značajnosti α, za testiranje hipoteze $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$, prema alternativnoj hipotezi $\mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2$, određeno je nejednakostima

$$v \leq F_{rs}^{-1}\left(\frac{\alpha}{2}\right), \quad v \geq F_{rs}^{-1}\left(1-\frac{\alpha}{2}\right), \quad r = m-1, \; s = n-1. \tag{103}$$



Slika 26. Skica kritičnog područja i krivulje razdiobe test-statistike V

Za ilustraciju opisanog testa promotrimo ovaj zadatak: U proizvodnji betona upotrebljavaju se dva tipa mješalica A i B. Mjerene su odgovarajuće tlačne čvrstoće betona i pritom su dobiveni ovi rezultati:

tip A	20,3	22,9	21,5	22,0	22,3	23,3	20,8	22,8
tip B	18,1	20,7	16,5	19,0	18,5	18,2.		

Postoji li značajna razlika u varijanci tlačne čvrstoće betona za jedan i drugi tip mješalice?

Ovdje je $m = 8$ i $n = 6$, pa uzme li se $\alpha = 0{,}10$, primjenom tabl. VII. u Dodatku, treba naći $F_{7,5}^{-1}(0{,}05)$ i $F_{7,5}^{-1}(0{,}95)$. Odmah se primjećuje da u tablici nema odgovarajućih vrijednosti za vjerojatnost 0,05, već se nalaze samo vrijednosti za vjerojatnosti 0,95 i 0,99. Međutim, zbog očigledne simetrije uloga uzoračkih varijanci S_x^2 i S_y^2 u (100), odmah se zaključuje da

$$\frac{S_y^2}{S_x^2} = \frac{1}{V} \sim \mathrm{F}(n-1, m-1). \tag{104}$$

To nas upućuje da općenito vrijedi

$$V \sim \mathrm{F}(r,s) \implies \frac{1}{V} \sim \mathrm{F}(s,r), \quad r,s \in \mathbf{N},$$

iz čega proizlazi da za svaki $c > 0$ vrijedi

$$F_{rs}(c) = 1 - F_{sr}\left(\frac{1}{c}\right) = \alpha,$$

odnosno

$$F_{rs}^{-1}(\alpha) = \frac{1}{F_{sr}^{-1}(1-\alpha)}. \tag{105}$$

Na temelju (105) vidi se da je $F_{7,5}^{-1}(0{,}05) = \dfrac{1}{F_{5,7}^{-1}(0{,}95)} = \dfrac{1}{3{,}97} = 0{,}25$, dok je $F_{7,5}^{-1}(0{,}95) = 4{,}48$, pa iz (103) proizlazi da je kritično područje $[0;0{,}25] \cup [4{,}48;\infty\rangle$.

Na temelju izmjerenih podataka nalazi se da je

$$\overline{x} = 21{,}0, \quad \overline{y} = 18{,}2, \quad s_x^2 = 2{,}23, \quad s_y^2 = 1{,}87,$$

pa iz (101) proizlazi da je $v = 1{,}19$, iz čega se vidi da dobivena vrijednost test-statistike V ne pada u kritično područje testa, što upućuje na zaključak da ne postoji značajna razlika u varijancama tlačne čvrstoće betona proizvedenog na mješalicama tipa A i tipa B.

8. Testovi o koeficijentu korelacije

U V.3. opisan je teorijski model za opisivanje praktične situacije pri mjerenju dviju slučajnih varijabli X i Y, koji se zove dvodimenzionalna normalna razdioba $\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, gdje parametar ρ $(-1 < \rho < 1)$ ima značenje koeficijenta korelacije slučajnih varijabli X i Y. Ako je $\rho = 0$, onda su X i Y nezavisne slučajne varijable, pa se prirodno nameće zadaća da se na temelju n $(n > 2)$ mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ slučajnog vektora (X, Y) donese odluka o zavisnosti ili nezavisnosti slučajnih varijabli X i Y. Ta se zadaća može formulirati kao testiranje nul-hipoteze $\mathrm{H}_0 : \rho = 0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \rho \neq 0$.

Odmah se nameće ideja da se ML-procjenitelj

$$\widehat{P} = \frac{\sum\limits_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2 \sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{106}$$

nepoznatog parametra ρ (v. VI.4) iskoristi pri definiranju pripadnoga kritičnog područja zadane razine značajnosti α. Da bi se to moglo učiniti trebalo bi poznavati razdiobu vjerojatnosti procjenitelja $\widehat{P}$, uz pretpostavku da je hipoteza $\mathrm{H}_0 : \rho = 0$ istinita. Ne može se, doduše, dobiti jednostavna vjerojatnosna razdioba neposredno za $\widehat{P}$, ali se pokazuje (v. [19]) da statistika

$$T = \frac{\widehat{P}}{\sqrt{1-\widehat{P}^2}}\sqrt{n-2} \sim \mathrm{t}(n-2). \tag{107}$$

To omogućuje da se slučajna varijabla T uzme kao test-statistika, čija će vrijednost

(108) $$\tau = \frac{\hat{\rho}}{\sqrt{1-\hat{\rho}^2}}\sqrt{n-2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2-\left[\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\right]^2}}\cdot\sqrt{n-2}$$

biti odlučujuća za donošenje odluke pri testiranju hipoteze H_0. Iz zahtjeva da test ima razinu značajnosti α, tj. iz

$$P_0(|T| > c_0) = \alpha,$$

proizlazi

$$c_0 = G_{n-2}^{-1}\left(1-\frac{\alpha}{2}\right),$$

gdje je G_{n-2}^{-1} inverzna funkcija od f.r.v. Studentove razdiobe $t(n-2)$ sa $n-2$ stupnja slobode.

Treba li, dakle, konstruirati test razine značajnosti α za testiranje hipoteze $H_0 : \rho = 0$, prema alternativnoj hipotezi $H_1 : \rho \neq 0$, pripadno kritično područje određeno je uvjetom

(109) $$|\tau| \geq G_{n-2}^{-1}\left(1-\frac{\alpha}{2}\right).$$

Uređene parove iz 1. primjera u III.1. možemo shvatiti kao mjerenja slučajnog vektora (X, Y), gdje X označuje ocjenu iz matematike, a Y ocjenu iz fizike, pa hipotezu $H_0 : \rho = 0$ možemo interpretirati kao hipotezu da nema značajne korelacije između ocjena iz matematike i fizike u određenoj učeničkoj populaciji. Mjerenja iz spomenutog primjera tretirat ćemo kao vrijednost slučajnog uzorka veličine $n = 30$, na temelju čega se dobiva vrijednost uzoračkog koeficijenta korelacije $\hat{\rho} = 0{,}75$, tako da odgovarajuća vrijednost test-statistike T iz (107) iznosi $\tau = 6{,}00$. Za $\alpha = 0{,}05$ dobiva se $G_{28}^{-1}(0{,}975) = 2{,}05$ (v. tabl. V. u Dodatku), pa se vidi da dobivena vrijednost test-statistike pada u kritično područje $\langle -\infty; -2{,}05] \cup [2{,}05; \infty\rangle$, što znači da treba odbaciti hipotezu $H_0 : \rho = 0$ i prihvatiti alternativnu hipotezu $H_1 : \rho \neq 0$.

Kao što se moglo i očekivati, vrijednost $\hat{\rho} = 0{,}75$ uzoračkog koeficijenta korelacije indicira da hipotezu o nekoreliranosti (nezavisnosti) ocjena iz matematike i fizike treba odbaciti. Tako visoka vrijednost uzoračkog koeficijenta korelacije potiče nas, dapače, da vjerujemo u postojanje jake koreliranosti između ocjena iz matematike i fizike. Stoga se odmah nameće zadatak da se konstruira test za testiranje hipoteze o pretpostavljenoj vrijednosti ρ_0 koeficijenta korelacije.

Pretpostavimo, dakle, da uz već navedene opće pretpostavke treba konstruirati test razine značajnosti α za testiranje hipoteze $H_0 : \rho = \rho_0$ $(-1 < \rho_0 < 1)$, prema alternativnoj hipotezi $H_1 : \rho \neq \rho_0$.

Zadatak se može vrlo jednostavno riješiti za velike n, jer će se iskoristiti činjenica da, u uvjetima istinitosti hipoteze H_0, slučajna varijabla

(110) $$W = \frac{1}{2}\ln\frac{1+\hat{P}}{1-\hat{P}}$$

ima asimptotski normalnu razdiobu s očekivanjem $\mu = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}$ i varijancom $\sigma^2 = \frac{1}{n-3}$. To omogućuje da se kritično područje razine značajnosti α definira pomoću test-statistike

(111) $$Z = \frac{W-\mu}{\sigma} = \frac{\sqrt{n-3}}{2}\ln\frac{(1+\hat{P})(1-\rho_0)}{(1-\hat{P})(1+\rho_0)} \sim N(0,1).$$

Iz $P_0(|Z| > c_0) = \alpha$ proizlazi $c_0 = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$, tako da je pripadno kritično područje C_0 određeno uvjetom

(112) $$|z| \geq \Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

gdje je

(113) $$z = \frac{\sqrt{n-3}}{2}\ln\frac{(1+\hat{\rho})(1-\rho_0)}{(1-\hat{\rho})(1+\rho_0)}$$

vrijednost test-statistike Z.

Uzme li se kao alternativna hipoteza $H_1 : \rho > \rho_0$, pripadno kritično područje bit će određeno nejednakošću (v. zad. 17)

(114) $$z \geq \Phi^{-1}(1-\alpha),$$

a ako je $H_1 : \rho < \rho_0$, pripadno kritično područje definirano je nejednakošću

(115) $$z \leq \Phi^{-1}(\alpha).$$

Na temelju podataka iz već spomenutog 1. primjera u III.1. mogla bi se testirati hipoteza $H_0 : \rho = 0{,}5$, prema alternativnoj hipotezi $H_1 : \rho < 0{,}5$. Test-statistika Z ima vrijednost $z = \frac{\sqrt{30-3}}{2}\ln\frac{(1+0{,}75)(1-0{,}5)}{(1-0{,}75)(1+0{,}5)} = 2{,}20$. Uzme li se $\alpha = 0{,}05$, dobiva se $\Phi^{-1}(0{,}05) = -1{,}65$, pa se, na temelju (115), zaključuje da dobivena vrijednost test-statistike ne pada u kritično područje $\langle -\infty; -1{,}65\rangle$, što znači da hipotezu H_0, tj. slutnju da su ocjene iz matematike i fizike značajno pozitivno korelirane, treba prihvatiti.

Zadaci

1. Slučajnim uzorkom veličine $n = 25$ želi se provjeriti tvrdnja dobavljača da su pakovanja soli, deklarirana kao pet kilogramska, korektna. Pretpostavlja se, zapravo, da je težina jednog pakovanja slučajna varijabla $X \sim \mathrm{N}(\mu, 1)$ i da je proces pakiranja, sa stajališta kupca, korektan ako je $\mu \geq 5$.

a) Formulirajte nul-hipotezu i alternativnu hipotezu.

b) Napišite formulu i skicirajte graf odgovarajuće idealne funkcije snage testa.

c) Uzmite kao kritično područje za nul-hipotezu skup $C = \{(x_1, \ldots, x_{25}) \in \mathbf{R}^{25} : \bar{x} < 5\}$, odredite pripadnu funkciju snage, skicirajte njen graf i nađite pripadnu razinu značajnosti testa.

d) Za kritično područje oblika $C = \{(x_1, \ldots, x_{25}) \in \mathbf{R}^{25} : \bar{x} < c\}$ odredite konstantu c tako da razina značajnosti testa iznosi $\alpha = 0{,}1$. Nađite pripadnu funkciju snage i skicirajte njen graf.

e) Odredite veličinu uzorka n i konstantu c u kritičnom području oblika $C = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : \bar{x} < c\}$ tako da za pripadnu funkciju snage K vrijedi $K(5) = 0{,}1$ i $K(5{,}2) = 0{,}9$. Skicirajte graf dobivene funkcije snage.

f) Kolika je vjerojatnost da će se, primjenom testa iz c), prihvatiti nul-hipoteza kada nepoznati parametar μ ima vrijednost 6, a kolika kada ima vrijednost 4? Kolike su spomenute vjerojatnosti u testu iz d), a kolike u testu iz e)?

2. Primalac velike pošiljke istovrsnih proizvoda želi provjeriti tvrdnju proizvođača da pošiljka ne sadrži više od 10% neispravnih proizvoda. U tu svrhu slučajno uzima $n = 10$ proizvoda iz pošiljke i utvrđuje broj X neispravnih proizvoda među njima. Pretpostavlja se da $X \sim \mathrm{B}(n, p)$ i da je, zapravo, tvrdnja proizvođača $p \leq 0{,}1$.

a) Opišite skup dopuštenih vrijednosti nepoznatog parametra p i formulirajte nul-hipotezu i alternativnu hipotezu.

b) Uzmite kao kritično područje skup $C = \{(x_1, \ldots, x_{10}) \in \mathbf{R}^{10} : \sum_{i=1}^{10} x_i \geq 1\}$, pa odredite pripadnu operativnu karakteristiku testa, skicirajte njen graf i nađite razinu značajnosti.

3. Postavlja se hipoteza da je godišnji broj pojava tuče (gràda) na određenoj lokaciji slučajna varijabla X Poissonove razdiobe parametra $\lambda = 0{,}5$. Na temelju podataka iz prethodnih $n = 12$ godina želi se testirati postavljena hipoteza $\mathrm{H}_0 : \lambda = 0{,}5$, prema alternativnoj hipotezi $\mathrm{H}_1 : \lambda > 0{,}5$.

a) Napišite formulu i skicirajte graf odgovarajuće idealne operativne karakteristike.

b) Kako glasi operativna karakteristika testa kojemu pripada kritično područje $C = \{(x_1, \ldots, x_{12}) \in \mathbf{R}^{12} : x_1 + \ldots + x_{12} < 2\}$?

c) Skicirajte graf operativne karakteristike testa iz b) i odredite pripadnu razinu značajnosti testa.

d) Kolika je vjerojatnost da će se odbaciti hipoteza H_0, primjenom testa iz b), kada je stvarna vrijednost nepoznatog parametra $\lambda = 1$, a kolika kada je $\lambda = 0{,}6$.

4. Pretpostavlja se da slučajna varijabla $X \sim \mathrm{U}(0, t)$ $(t > 0)$, gdje je t nepoznati parametar za kojega se postavlja nul-hipoteza $\mathrm{H}_0 : t = 1$, prema alternativnoj hipotezi $\mathrm{H}_1 : t \neq 1$. Na temelju $n = 4$ mjerenja slučajne varijable X dobivene su vrijednosti x_1, x_2, x_3 i x_4.

a) Uzmite skup $C = \{(x_1, x_2, x_3, x_4) \in \mathbf{R}^4 : |\max(x_1, x_2, x_3, x_4) - 1| > 0{,}5\}$ kao kritično područje testa, odredite pripadnu funkciju snage i razinu značajnosti.

b) Kolika je vjerojatnost da se prihvati hipoteza H_0 kada je stvarna vrijednost nepoznatog parametra $t = 1{,}2$, a kolika za $t = 0{,}8$?

5. Načinjena su dva mjerenja x_1, x_2 slučajne varijable $X \sim \mathrm{Ex}(\alpha)$ $(\alpha > 0)$. Postavlja se nul-hipoteza $\mathrm{H}_0 : \alpha = 1$, prema alternativnoj hipotezi $\mathrm{H}_1 : \alpha = 2$, i definira kritično područje $C = \{(x_1, x_2) \in \mathbf{R}^2 : x_1 + x_2 > 4\}$. Odredite:

a) operativnu karakteristiku,

b) razinu značajnosti,

c) vjerojatnost da se H_0 prihvati, kada stavarno nije istinita,

d) vjerojatnost da se H_0 odbaci, kada je stvarno istinita.

6. Primjenom Neyman-Pearsonove leme nađite najbolje kritično područje za testiranje jednostavne hipoteze $\mathrm{H}_0 : \lambda = 1$, prema jednostavnoj alternativnoj hipotezi $\mathrm{H}_1 : \lambda = 1{,}5$, uz pretpostavku da mjerenja potječu od slučajne varijable $X \sim \mathrm{Po}(\lambda)$ i da je $n = 10$ i $\alpha = 0{,}01$. Kolike su pripadne pogreške prve i druge vrste?

7. Pretpostavlja se da slučajna varijabla X ima beta-razdiobu s parametrom $\alpha = 1$ i nepoznatim parametrom $\beta > 0$. Dokažite da najbolje kritično područje za testiranje jednostavne hipoteze $\mathrm{H}_0 : \beta = 1$, prema jednostavnoj alternativnoj hipotezi $\mathrm{H}_1 : \beta = 2$, ima oblik

$$C = \{(x_1, \ldots, x_n) \in \mathbf{R}^n : x_1 x_2 \cdot \ldots \cdot x_n > c\},$$

gdje je $c > 0$ određeni realni broj.

8. Parametri μ i σ^2 normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$ tretiraju se kao komponente vektorskog parametra $\boldsymbol{t} = (\mu, \sigma^2)$ sa skupom dopuštenih vrijednosti $\boldsymbol{\Theta} = \{(\mu, \sigma^2) : \mu \in \mathbf{R},\ \sigma > 0\}$. Dokažite da je

$$C = \left\{ (x_1, \ldots, x_n) \in \mathbf{R}^n : \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i > c \right\}$$

najbolje kritično područje za testiranje jednostavne hipoteze $\mathrm{H}_0 : \boldsymbol{t} = \boldsymbol{t}_0 = (0, 1)$ $(\mu = 0,\ \sigma^2 = 1)$, prema jednostavnoj alternativnoj hipotezi $\mathrm{H}_1 : \boldsymbol{t} = \boldsymbol{t}_1 = (0, 4)$ $(\mu = 0,\ \sigma^2 = 4)$.

9. Konstruirajte jednoliko najsnažniji test, zadane razine značajnosti α, za testiranje jednostavne hipoteze $\mathrm{H}_0 : \lambda = 0{,}5$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \lambda > 0{,}5$, pri čemu se pretpostavlja da mjerenja $x_1, \ldots, x_n$ potječu od

slučajne varijable $X \sim \mathrm{Po}(\lambda)$. Uzmite zatim $n = 12$ i $\alpha = 0{,}05$ te usporedite dobiveni rezultat s onim iz 3. primjera.

10. Odredite kritično područje jednoliko najsnažnijeg testa, razine značajnosti α, za testiranje jednostavne hipoteze $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$, prema složenoj alternativnoj hipotezi $\mathrm{H}_1 : \sigma^2 < \sigma_0^2$ ($\sigma_0 > 0$ je zadani broj), pri čemu se pretpostavlja da mjerenja $x_1, \ldots, x_n$ potječu od slučajne varijable $X \sim \mathrm{N}(0, \sigma^2)$. Uzmite $\mathrm{H}_1 : \sigma^2 \neq \sigma_0^2$ kao alternativnu hipotezu i pokažite da ne postoji jednoliko najsnažniji test.

11. Dokažite da LR-testovima (a)-(d) u VIII.6. pripadaju kritična područja (a_1)-(d_1), pri čemu se pretpostavlja da su ispunjene pretpostavke iz 5. primjera.

12. Dokažite da LR-testovima (a)-(d) u VIII.6. pripadaju kritična područja (a_2)-(d_2), pri čemu se pretpostavlja da su ispunjene pretpostavke iz 6. primjera.

13. Dokažite da se metodom primjene intervala povjerenja dobiva isto kritično područje kao i metodom omjera vjerodostojnosti za probleme iz 5. i 6. primjera.

14. Mjere se dvije nezavisne slučajne varijable X i Y, kojima pripadaju normalne razdiobe s varijancama σ_1^2 i σ_2^2. Načinjeno je m mjerenja $x_1, \ldots, x_m$ slučajne varijable X i n mjerenja $y_1, \ldots, y_n$ slučajne varijable Y. Konstruirajte test razine značajnosti α za testiranje hipoteze $\mathrm{H}_0 : \mu_1 - \mu_2 = d_0$ (d_0 je zadani realni broj), prema alternativnoj hipotezi $\mathrm{H}_1 : \mu_1 - \mu_2 \neq d_0$, uz pretpostavku:

a) da su varijance σ_1^2 i σ_2^2 poznate,

b) da je $\sigma_1^2 = \sigma_2^2 = \sigma^2$ nepoznato.

Uputa: Primijenite rezultate zad. 7. iz VII. poglavlja i metodu intervala povjerenja.

15. Odredite uvjete kojima se definira kritično područje testa razine značajnosti α, pri testiranju hipoteze $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$, prema alternativnoj hipotezi $\mathrm{H}_1 : \sigma_1^2 < \sigma_2^2$ ($\sigma_1^2 > \sigma_2^2$), uz pretpostavke iz 9. primjera.

16. Odredite uvjete kojima se definira kritično područje testa razine značajnosti α, pri testiranju hipoteze $\mathrm{H}_0 : \rho = 0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \rho < 0$ ($\rho > 0$), uz primjenu test-statistike $\hat{P}$ iz (106).

17. Odredite uvjete kojima se definira kritično područje testa, razine značajnosti α, pri testiranju hipoteze $\mathrm{H}_0 : \rho = \rho_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : \rho < \rho_0$ ($\rho > \rho_0$), uz primjenu test-statistike Z iz (111).

18. Primjenom rezultata iz VII.5. izvedite uvjete kojima se definira kritično područje testa, razine značajnosti α, pri testiranju:

a) $\mathrm{H}_0 : p = p_0$, prema $\mathrm{H}_1 : p \neq p_0$,

b) $\mathrm{H}_0 : p = p_0$, prema $\mathrm{H}_1 : p < p_0$,

c) $\mathrm{H}_0 : p = p_0$, prema $\mathrm{H}_1 : p > p_0$.

gdje je p nepoznata vjerojatnost promatranog događaja, p_0 ($0 < p_0 < 1$) zadani broj, a raspolaže se sa n mjerenja slučajne varijable $X \sim \mathrm{B}(1, p)$.

19. Na temelju podataka iz 3. primjera u I.2. (tabl. 3) testirajte, uz razinu značajnosti $\alpha = 0{,}05$, hipotezu da očekivani dnevni broj kvarova iznosi 10, prema alternativnoj hipotezi da je veći od 10. Testirajte i hipotezu da varijanca iznosi 9, prema alternativnoj hipotezi da je različita od 9.

20. Na temelju podataka iz 5. primjera u I.4. testirajte, uz razinu značajnosti $\alpha = 0{,}01$, hipotezu da je očekivana tlačna čvrstoća betona 30 MPa, prema alternativnoj hipotezi da je ona manja od 30 MPa. Testirajte i hipotezu da je varijanca 25 $(\mathrm{MPa})^2$, prema alternativnoj hipotezi da je ona veća od 25 $(\mathrm{MPa})^2$.

21. Smije li se na temelju podataka iz zad. 4. u I. poglavlju zaključiti da je očekivani broj dana bez oborina u mjesecu rujnu veći od 20?

22. Može li se u teoriji testiranja statističkih hipoteza naći oslonac za zaključak da podaci iz:

a) zad. 5. u I. poglavlju potkrepljuju tvrdnju da je dnevni broj prodanih cipela u promatranoj prodavaonici slučajna varijabla s očekivanjem $\mu = 100$ i standardnom devijacijom $\sigma = 10$,

b) zad. 6. u I. poglavlju potkrepljuju tvrdnju da je tjedni broj kvarova na strojevima promatranoga industrijskog pogona slučajna varijabla s očekivanjem $\mu = 1{,}5$ i varijancom $\sigma^2 = 0{,}25$?

23. Može li se na temelju podataka iz zad. 7. u I. poglavlju zaključiti da je očekivani broj telefonskih razgovora preko promatrane telefonske centrale u jednom satu jednak 25?

24. Na temelju podataka iz zad. 8. u I. poglavlju testirajte, uz razinu značajnosti $\alpha = 0{,}05$, hipotezu da promatrana čelična žica ima očekivanu čvrstoću od 300 MPa.

25. Uz pretpostavku da podaci iz zad. 13. u I. poglavlju potječu od normalne razdiobe, testirajte hipotezu, uz razinu značajnosti $\alpha = 0{,}1$, da se tlačna čvrstoća cementnih kocki rasipa s varijancom koja nije veća od 16 $(\mathrm{MPa})^2$.

26. Na temelju podataka iz 4. primjera u I.3. (tabl. 4) testirajte, uz razinu značajnosti $\alpha = 0{,}01$, hipotezu da vjerojatnost pojavljivanja slova A u tekstovima hrvatskog jezika iznosi 0,15, a vjerojatnost pojavljivanja slova B 0,02.

27. Može li se na temelju podataka iz zad. 2. u III. poglavlju zaključiti da nema značajne razlike u srednjim ocjenama iz matematike u završnom razredu srednje škole i na fakultetskom ispitu iz matematike?

28. Uz pretpostavku da podaci u 3. primjeru u III. poglavlju potječu od dvodimenzionalne normalne razdiobe, testirajte, uz razinu značajnosti $\alpha = 0{,}05$, hipotezu da su X i Y nekorelirane slučajne varijable. Je li prihvatljiva hipoteza da su one pozitivno ili da su negativno korelirane slučajne varijable?

Pregled važnijih parametarskih testova

Osnovne pretpostavke		Nul-hipoteza	Alternativna hipoteza	Vrijednost test-statistike	Razdioba test-statistike	Opis kritičnog područja
mjerenja $x_1, \dots, x_n$ potječu od normalne razdiobe $N(\mu, \sigma^2)$	σ poznato	$\mu = \mu_0$	$\mu \neq \mu_0$	$z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}$	$N(0,1)$	$\lvert z\rvert \geq \Phi^{-1}(1 - \frac{\alpha}{2})$
			$\mu < \mu_0$			$z \leq \Phi^{-1}(\alpha)$
			$\mu > \mu_0$			$z \geq \Phi^{-1}(1 - \alpha)$
	σ nepoznato		$\mu \neq \mu_0$	$\tau = \frac{\overline{x} - \mu_0}{s}\sqrt{n}$	$t(n-1)$	$\lvert\tau\rvert \geq G_{n-1}^{-1}(1 - \frac{\alpha}{2})$
			$\mu < \mu_0$			$\tau \leq G_{n-1}^{-1}(\alpha)$
			$\mu > \mu_0$			$\tau \geq G_{n-1}^{-1}(1 - \alpha)$
	μ nepoznato	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$u = \frac{n-1}{\sigma_0^2}s^2$	$\chi^2(n-1)$	$u \leq H_{n-1}^{-1}(\frac{\alpha}{2})$ $u \geq H_{n-1}^{-1}(1 - \frac{\alpha}{2})$
			$\sigma^2 < \sigma_0^2$			$u \leq H_{n-1}^{-1}(\alpha)$
			$\sigma^2 > \sigma_0^2$			$u \geq H_{n-1}^{-1}(1 - \alpha)$
mjerenja potječu od $B(1, p)$, $n > 40$		$p = p_0$	$p \neq p_0$	$z = \frac{\overline{x} - p_0}{\sqrt{p_0(1 - p_0)}}\sqrt{n}$	$N(0,1)$	$\lvert z\rvert \geq \Phi^{-1}(1 - \frac{\alpha}{2})$
			$p < p_0$			$z \leq \Phi^{-1}(\alpha)$
			$p > p_0$			$z \geq \Phi^{-1}(1 - \alpha)$

Nastavak tablice

Osnovne pretpostavke		Nul-hipoteza	Alternativna hipoteza	Vrijednost test-statistike	Razdioba test-statistike	Opis kritičnog područja
mjerenja $x_1, \dots, x_m$ i $y_1, \dots, y_n$ potječu od nezavisnih normalnih razdioba $N(\mu_1, \sigma_1^2)$ i $N(\mu_2, \sigma_2^2)$	σ_1, σ_2 poznato	$\mu_1 - \mu_2 = d_0$	$\mu_1 - \mu_2 \neq d_0$	$z = \frac{\overline{x} - \overline{y} - d_0}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$	$N(0,1)$	$\lvert z\rvert \geq \Phi^{-1}(1 - \frac{\alpha}{2})$
			$\mu_1 - \mu_2 < d_0$			$z \leq \Phi^{-1}(\alpha)$
	$\sigma_1^2 = \sigma_2^2$ nepoznato		$\mu_1 - \mu_2 \neq d_0$	$\tau = \frac{\overline{x} - \overline{y} - d_0}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$, $s^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}$	$t(r)$ $r = m + n - 2$	$\lvert\tau\rvert \leq G_r^{-1}(1 - \frac{\alpha}{2})$
			$\mu_1 - \mu_2 < d_0$			$\tau \leq G_r^{-1}(\alpha)$
	μ_1, μ_2 nepoznati	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$v = \frac{s_x^2}{s_y^2}$	$F(r, s)$ $r = m - 1$ $s = n - 1$	$v \leq F_{rs}^{-1}(\frac{\alpha}{2})$, $v \geq F_{rs}^{-1}(1 - \frac{\alpha}{2})$
			$\sigma_1^2 < \sigma_2^2$			$v \geq F_{rs}^{-1}(1 - \alpha)$
mjerenja $(x_1, y_1), \dots$ $\dots, (x_n, y_n)$ potječu od $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \varrho)$, $n > 30$		$\varrho = 0$	$\varrho \neq 0$	$\tau = \frac{\widehat{\varrho}}{\sqrt{1 - \widehat{\varrho}^2}}\sqrt{n-2}$	$t(n-2)$	$\lvert\tau\rvert \geq G_{n-2}^{-1}(1 - \frac{\alpha}{2})$
			$\varrho < 0$			$\tau \leq G_{n-2}^{-1}(\alpha)$
			$\varrho > 0$			$\tau \geq G_{n-2}^{-1}(1 - \alpha)$
		$\varrho = \varrho_0$	$\varrho \neq \varrho_0$	$z = \frac{\sqrt{n-3}}{2}\ln\frac{(1 + \widehat{\varrho})(1 - \varrho_0)}{(1 - \widehat{\varrho})(1 + \varrho_0)}$	$N(0,1)$	$\lvert z\rvert \geq \Phi^{-1}(1 - \frac{\alpha}{2})$
			$\varrho < \varrho_0$			$z \leq \Phi^{-1}(\alpha)$
			$\varrho > \varrho_0$			$z \geq \Phi^{-1}(1 - \alpha)$

IX. Hikvadrat-test

1. Pearsonov teorem

Jedan od prvih testova, koji je 1900. godine predložio *K. Pearson*, jest čuveni hikvadrat-test. Matematički model na kojem se zasniva hikvadrat-test vrlo je jednostavan. Promatra se slučajni eksperiment (slučajna pojava) s konačnim skupom svih mogućih ishoda $A = \{a_1, \dots, a_r\}$ $(r \geq 2)$, tako da se načini n $(n \in \mathbf{N})$ nezavisnih ponavljanja toga eksperimenta i registrira frekvencija $\widehat{f}_j$, odnosno relativna frekvencija $\widehat{p}_j = \frac{1}{n}\widehat{f}_j$ $(j = 1, \dots, r)$, ishoda a_j. Pretpostavlja se da ishodu a_j pripada odgovarajuća vjerojatnost $p_j (p_j \geq 0,\ \sum_{j=1}^{r} p_j = 1)$. Vjerojatnosti $p_1, \dots, p_r$ obično su nepoznate, ali priroda promatrane pojave redovito upućuje na određenu hipotezu o njihovim vrijednostima, pa se odmah nameće zadatak da se utvrdi u kojoj mjeri dobiveni podaci (opažene frekvencije, odnosno relativne frekvencije) potkrepljuju ili opovrgavaju postavljenu hipotezu.

1. primjer

Igraća kocka bačena je $n = 100$ puta i pritom su registrirane relativne frekvencije brojeva (mogućih ishoda) iz skupa $A = \{1, 2, 3, 4, 5, 6\}$. Dobiveno je

$$\widehat{p}_1 = 0{,}22\,, \quad \widehat{p}_2 = 0{,}16\,, \quad \widehat{p}_3 = 0{,}22\,, \quad \widehat{p}_4 = 0{,}20\,, \quad \widehat{p}_5 = 0{,}08\,, \quad \widehat{p}_6 = 0{,}12\,.$$

Može li se, na temelju dobivenih podataka, zaključiti da je igraća kocka pravilna, ili pak treba zaključiti da postoje određene nepravilnosti u izradbi kocke? Egzaktnije postavljajući problem rekli bismo da treba testirati hipotezu

$$\mathrm{H}_0 \ : \ p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}\,,$$

prema alternativnoj hipotezi

$$\mathrm{H}_1 \ : \ \text{svi ishodi nemaju istu vjerojatnost } \frac{1}{6}.$$

Općenito problem se sastoji u tome da se konstruira test za testiranje nul-hipoteze $\mathrm{H}_0 : (p_1 = p_1^{(0)}, \dots, p_r = p_r^{(0)})$, gdje su $p_1^{(0)}, \dots, p_r^{(0)}$ unaprijed zadani brojevi $(p_j^{(0)} \geq 0,\ \sum_{i=1}^{r} p_j^{(0)} = 1)$, prema alternativnoj hipotezi H_1: (bar jedna od nepoznatih vjerojatnosti $p_1, \dots, p_r$ različita je od odgovarajuće pretpostavljene vrijednosti).

Ako su $a_1, \dots, a_r$ realni brojevi, onda se A može shvatiti kao skup vrijednosti diskretne slučajne varijable X, a $p_1, \dots, p_r$ kao pripadne vjerojatnosti, pa se hipoteza H_0 može interpretirati kao hipoteza da diskretnoj slučajnoj varijabli X pripada pretpostavljena razdioba vjerojatnosti $p_1, \dots, p_r$.

Odmah primjetimo da se problem može tretirati kao zadatak o testiranju parametarske hipoteze s vektorskim parametrom $\boldsymbol{p} = (p_1, \dots, p_r)$, pri čemu skup dopuštenih vrijednosti za vektorski parametar $\boldsymbol{p}$ glasi

$$\boldsymbol{\Theta} = \{\boldsymbol{p} \in \mathbf{R}^r : p_j > 0\,,\ \sum_{j=1}^{r} p_j = 1\}\,, \tag{1}$$

pa se nul-hipoteza može zapisati u obliku

$$\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^{(0)}\,, \tag{2}$$

a alternativna hipoteza

$$\mathrm{H}_1 : \boldsymbol{p} \neq \boldsymbol{p}^{(0)}\,. \tag{3}$$

Ideja vodilja pri konstrukciji testa, zadane razine značajnosti α, ista je kao i u VIII.7. To znači da treba poći od dobrog procjenitelja $\widehat{\boldsymbol{P}}$ za nepoznati parametar $\boldsymbol{p}$ i zatim definirati prikladnu test-statistiku, označimo je sa D, koja će indicirati razliku između vrijednosti $\widehat{\boldsymbol{p}}$ procjenitelja $\widehat{\boldsymbol{P}}$ i pretpostavljene vrijednosti $\boldsymbol{p}^{(0)}$. Dobije li se prevelika vrijednost d test-statistike D, hipoteza H_0 će se odbaciti. Da bi se odredilo pripadno kritično područje, uza zadanu razinu značajnosti α, nužno je poznavati vjerojatnosnu razdiobu test-statistike D, pri pretpostavci da je hipoteza H_0 stvarno istinita. Nastoji se, dakako, naći ona test-statistika koja će imati što jednostavniju razdiobu vjerojatnosti. Iduće razmatranje dovodi nas do takve test-statistike.

Uoči li se ishod $a_j \in A$, pripadna relativna frekvencija $\widehat{p}_j = \frac{1}{n}\widehat{f}_j$ može se tretirati kao vrijednost statistike $\widehat{P}_j = \frac{1}{n}\widehat{F}_j$ (uzoračka relativna frekvencija), za koju se zna da je ML-procjenitelj nepoznatog parametra p_j. Ako je hipoteza H_0 stvarno istinita, može se očekivati da ponderirana suma kvadrata odstupanja

$$d = \sum_{j=1}^{r} \lambda_j (\widehat{p}_j - p_j^{(0)})^2 \tag{4}$$

neće biti velika. Dobije li se $d \geq c_0$, hipoteza H_0 će se odbaciti, pri čemu je konstanta $c_0 > 0$ određena tako da test ima razinu značajnosti α, tj. da vrijedi

$$P_0(D \geq c_0) = \alpha\,, \tag{5}$$

gdje je

$$D = \sum_{j=1}^{r} \lambda_j (\widehat{P}_j - p_j^{(0)})^2\,. \tag{6}$$

Oznaka P_0 u (5) označuje vjerojatnost navedenog događaja uz pretpostavku da je istinita hipoteza $\mathrm{H}_0 : \boldsymbol{p} = \boldsymbol{p}^{(0)}$.

Da bi se odredila vjerojatnosna razdioba test-statistike D iz (6), rezonira se ovako: $\widehat{P}_j$ je nepristran, konzistentan i asimptotski normalan procjenitelj za nepoznati parametar p_j, pa ako je istinita hipoteza H_0, onda je $p_j = p_j^{(0)}$ i slučajna varijabla $Z_j = \dfrac{\widehat{P}_j - p_j^{(0)}}{\sqrt{p_j^{(0)}(1-p_j^{(0)})}}\sqrt{n}$, za velike n ($n \to \infty$), približno ima standardnu normalnu razdiobu $N(0,1)$. Uzme li se $\lambda_j = \dfrac{n}{p_j^{(0)}(1-p_j^{(0)})}$, odmah se vidi da se (6) može pisati kao $D = \sum\limits_{j=1}^{r} Z_j^2$, što pokazuje da je slučajna varijabla D izražena kao zbroj kvadrata standardnih normalnih slučajnih varijabli. One, nažalost, nisu i nezavisne, jer između slučajnih varijabli $\widehat{P}_1, \ldots, \widehat{P}_r$ postoji linearna zavisnost izražena jednadžbom $\sum\limits_{j=1}^{r} \widehat{P}_j = 1$, što onemogućuje zaključak da $D \sim \mathcal{X}^2(r)$, što bi inače proizašlo iz rezultata navedenog u točki 5. u V.6. Može se, međutim, dokazati (v. [38]) da vrijedi tzv. *Pearsonov teorem*:

> U uvjetima istinitosti hipoteze $H_0 : p = p^{(0)}$, za velike n, test-statistika
>
> (7) $$D = \sum_{j=1}^{r} \frac{n}{p_j^{(0)}} (\widehat{P}_j - p_j^{(0)})^2 = \sum_{j=1}^{r} \frac{(\widehat{F}_j - f_j^{(0)})^2}{f_j^{(0)}},$$
>
> gdje je $f_j^{(0)} = np_j^{(0)}$, približno ima hikvadrat-razdiobu sa $r-1$ stupnjeva slobode.

Prema tome, da bi se odredilo kritično područje razine značajnosti α za testiranje nul-hipoteze (2), prema alternativnoj hipotezi (3), jednadžba (5) će se pomoću f.r.v. H_{r-1} hikvadrat-razdiobe $\mathcal{X}^2(r-1)$ zapisati

$$1 - P_0(D < c_0) = 1 - H_{r-1}(c_0) = \alpha,$$

iz čega odmah slijedi

(8) $$c_0 = H_{r-1}^{-1}(1-\alpha),$$

tako da je traženo kritično područje opisano uvjetom

(9) $$d = \sum_{j=1}^{r} \frac{(\widehat{f}_j - f_j^{(0)})^2}{f_j^{(0)}} \geq H_{r-1}^{-1}(1-\alpha),$$

gdje je $\widehat{f}_j$ izmjerena (opažena) frekvencija, a $f_j^{(0)} = np_j^{(0)}$ očekivana (teorijska) frekvencija ishoda a_j u nizu od n nezavisnih ponavljanja danoga slučajnog eksperimenta.

Za podatke iz 1. primjera, gdje je $r = 6$ i $f_j^{(0)} = 100 \cdot \frac{1}{6} \approx 16{,}67$ ($j = 1,2,3,4,5,6$) dobiva se $d \approx 10{,}91$. Uzme li se $\alpha = 0{,}05$, primjenom tabl. VI. u Dodatku, nalazi se da je $H_5^{-1}(0{,}95) = 11{,}1$, pa se vidi da dobivena vrijednost test-statistike ne pada u kritično područje $[11{,}1\ ;\ \infty\rangle$. To znači da hipotezu H_0 treba prihvatiti, odnosno da opaženi podaci ne upućuju na bitne nepravilnosti u ponašanju igraće kocke.

2. Fisherov teorem

Jedna od najvažnijih primjena hikvadrat-testa jest provjera hipoteze o tipu vjerojatnosne razdiobe. Da bi se lakše i bolje shvatila primjena hikvadrat-testa pri testiranju hipoteze o tipu vjerojatnosne razdiobe iz koje potječu izmjereni podaci, razmotrit će se najprije jedan ilustrativan primjer.

2. primjer

U 3. primjeru iz I.2. navedena je tablica frekvencija (tabl. 3) kvarova na strojevima određenoga industrijskog pogona. Pojednostavnjeno, ta tablica izgleda ovako:

Tablica 1.

Dnevni broj kvarova (j)	5 ili manje	6	7	8	9	10	11	12	13	14	15	16 ili više
Frekvencija (f_j)	11	13	20	18	27	34	28	12	13	10	7	7

Postavlja se pitanje može li se smatrati da navedeni podaci potječu od slučajne varijable X (dnevni broj kvarova) kojoj pripada Poissonova razdioba $Po(\lambda)$, pri čemu je parametar λ zasada još neodređen. Kada bi λ bio poznat, onda bi vrijednosti $j \in \{0,1,2,\ldots\}$ pripadala, u Poissonovoj razdiobi, vjerojatnost $p_j(\lambda) = \dfrac{\lambda^j}{j!}\exp(-\lambda)$. Tada bi i veličina d, koja pokazuje razliku između empirijskih (opaženih) frekvencija $\widehat{f}_j$ i teorijskih (očekivanih) frekvencija

$$f_j(\lambda) = np_j(\lambda) = n \cdot \frac{\lambda^j}{j!}\exp(-\lambda)$$

ovisila o λ, tako da ćemo pisati

(10) $$d(\lambda) = \sum_{j=0}^{\infty} \frac{[\widehat{f}_j - np_j(\lambda)]^2}{np_j(\lambda)} = \sum_{j=0}^{\infty} \frac{[\widehat{f}_j - f_j(\lambda)]^2}{f_j(\lambda)}.$$

Da bi se našla ona Poissonova razdioba $Po(\lambda)$ koja je najbolje prilagođena opaženim frekvencijama $\widehat{f}_j$, čini se razumnim odrediti parametar λ tako da $d(\lambda)$ bude minimalno, tj. naći takvo $\widehat{\lambda}$ da vrijedi

(11) $$\min_{\lambda \geq 0} d(\lambda) = d(\widehat{\lambda}).$$

Deriviranjem jednadžbe (10) po λ i sređivanjem dobiva se

(12) $$d'(\lambda) = -\frac{1}{n} \sum_{j=0}^{\infty} \frac{\widehat{f}_j^2\, p_j'(\lambda)}{[p_j(\lambda)]^2}.$$

Da bi se riješila jednadžba $d'(\lambda) = 0$, po λ, uzet će se u obzir da, za velike n i $\lambda \approx \widehat{\lambda}$, približno vrijedi $n\,p_j(\lambda) \approx \widehat{f}_j$, pa se dobiva

$$\sum_{j=0}^{\infty} \frac{\widehat{f}_j\, p_j'(\lambda)}{p_j(\lambda)} = \sum_{j=0}^{\infty} \widehat{f}_j \left(\frac{j}{\lambda} - 1\right) = 0\,,$$

iz čega proizlazi da je

(13) $$\lambda = \widehat{\lambda} = \frac{1}{n}\sum_{j=0}^{\infty} j \cdot \widehat{f}_j\,.$$

Vidi se da je $\widehat{\lambda}$, zapravo, aritmetička sredina izmjerenih podataka pa se može reći da je Poissonova razdioba s parametrom $\lambda = \widehat{\lambda} = \overline{x}$ nabolje prilagođena opaženim (izmjerenim) frekvencijama $\widehat{f}_j$. Primijetimo da je $\overline{x}$ i vrijednost ML-procjenitelja za nepoznati parametar Poissonove razdiobe (v. 5. primjer u VI.4).

Na temelju danih podataka može se izračunati (ne baš posve precizno, jer su u navedenoj tablici grupirani podaci za $j \le 5$ i $j \ge 16$)

$$\overline{x} = \widehat{\lambda} = \frac{1}{200}\,(5 \cdot 11 + 6 \cdot 13 + \ldots + 15 \cdot 7 + 16 \cdot 7) = 9{,}89 \approx 10\,.$$

Primjenom Poissonove razdiobe Po(10) izračunajmo teorijske vjerojatnosti

$$\begin{aligned}
\overline{p}_5(10) &= p_0(10) + p_1(10) + p_2(10) + p_3(10) + p_4(10) + p_5(10) = 0{,}067\\
\overline{p}_6(10) &= p_6(10) = 0{,}063\\
&\vdots\\
\overline{p}_{15}(10) &= p_{15}(10) = 0{,}034\\
\overline{p}_{16}(10) &= p_{16}(10) + p_{17}(10) + \cdots = 0{,}049\,,
\end{aligned}$$

kojima odgovaraju teorijske frekvencije

$$\begin{aligned}
\overline{f}_5(10) &= 200 \cdot 0{,}067 = 13{,}4\\
\overline{f}_6(10) &= 200 \cdot 0{,}063 = 12{,}6\\
&\vdots\\
\overline{f}_{15}(10) &= 200 \cdot 0{,}034 = 6{,}8\\
\overline{f}_{16}(10) &= 200 \cdot 0{,}049 = 9{,}8\,.
\end{aligned}$$

To omogućuje da se izračuna

(14) $$d(\widehat{\lambda}) = d(10) = \sum_{j=5}^{16} \frac{[\widehat{f}_j - \overline{f}_j(10)]^2}{\overline{f}_j(10)} = 9{,}57\,,$$

što na određeni način upućuje na globalnu razliku između empirijskih i teorijskih frekvencija u promatranom primjeru, pa se odmah postavlja pitanje da li je ta razlika dovoljno velika da se odbaci hipoteza da podaci potječu od Poissonove razdiobe

Po(10). Odgovor na postavljeno pitanje dobit ćemo onda, kada uspijemo utvrditi da je $d(\widehat{\lambda})$ vrijednost određene test-statistike s poznatom razdiobom vjerojatnosti.

Najprije primijetimo da teorijske frekvencije $\overline{f}_j(\widehat{\lambda}) = f_j(10)$ $(j = 5, \ldots, 16)$ ovise o vrijednosti $\widehat{\lambda} = 10$ ML-procjenitelja $\widehat{\Lambda}$ za nepoznati parametar λ Poissonove razdiobe, pa se stoga $\overline{f}_j(\widehat{\lambda})$ treba shvatiti kao vrijednost statistike $\overline{f}_j(\widehat{\Lambda})$. Stoga se i $d(\widehat{\lambda})$ treba shvatiti kao vrijednost slučajne varijable

(15) $$D(\widehat{\Lambda}) = \sum_{j=5}^{16} \frac{[\widehat{F}_j - \overline{f}_j(\widehat{\Lambda})]^2}{\overline{f}_j(\widehat{\Lambda})}\,.$$

Prema tome, ako se veličina $d(\widehat{\lambda})$ iz (14) želi iskoristiti kao kriterij za donošenje odluke pri testiranju hipoteze H_0 da zadani podaci potječu od Poissonove razdiobe, uz dani rizik α da će se odbaciti istinita hipoteza, nužno je poznavati razdiobu vjerojatnosti slučajne varijable $D(\widehat{\Lambda})$. U promatranom je primjeru $r = 12$, pa bi se na temelju ranijih razmatranja moglo pomisliti da za velike n približno vrijedi $D(\widehat{\Lambda}) \sim \chi^2(r-1)$. Međutim, usporedbom formula (7) i (15), odmah se vidi da je $D(\widehat{\Lambda})$ ipak različita slučajna varijabla od D, pa nema čvrstih razloga za vjerovanje da imaju jednake vjerojatnosne razdiobe. Dokazuje se (v. [6]), dapače, da vrijedi tzv. *Fisherov teorem*:

Neka vjerojatnosti mogućih ishoda $a_j \in A\,(j = 1, \ldots, r)$ ovise o parametru (može biti i vektorski) $t \in \Theta$, tj. $p_j = p_j(t) \ge 0$ $\left(\sum_{j=1}^{r} p_j(t) = 1,\ \forall t \in \Theta\right)$, i neka se, pri n nezavisnih ponavljanja toga slučajnog eksperimenta, dobije frekvencija $\widehat{f}_j$ ishoda a_j. Neka je, nadalje, $\widehat{T}$ određeni procjenitelj za nepoznati parametar t, čija se vrijednost $\widehat{t}$ određuje uvjetom

(16) $$d(\widehat{t}) = \min_{t\in\Theta} d(t)\,,$$

gdje je

$$d(t) = \sum_{j=1}^{r} \frac{[\widehat{f}_j - f_j(t)]^2}{f_j(t)}\,, \qquad f_j(t) = n\,p_j(t)\,, \qquad j = 1, \ldots, r\,.$$

Tada, za velike n, slučajnoj varijabli

(17) $$D(\widehat{T}) = \sum_{j=1}^{r} \frac{[\widehat{F}_j - f_j(\widehat{T})]^2}{f_j(\widehat{T})}\,,$$

gdje je $\widehat{F}_j$ statistika s vrijednostima $\widehat{f}_j$ (uzoračka frekvencija), pripada hikvadrat-razdioba sa $r - v - 1$ stupnjeva slobode, pri čemu v označuje dimenziju parametra t.

Procjenitelj $\widehat{T}$, dobiven na upravo opisani način, zove se *minimalni hikvadratni procjenitelj*, koji se u nekim slučajevima podudara sa ML-procjeniteljem.

Fisherov teorem općenito omogućuje da se konstruira test zadane razine značajnosti α, za testiranje hipoteze H_0: (podaci potječu od diskretne vjerojatnosne razdiobe ovisne o parametru t), prema alternativnoj hipotezi H_1: (podaci ne potječu od pretpostavljene vjerojatnosne razdiobe). Iz činjenice da test-statistika $D(\widehat{T}) \sim X^2(r-v-1)$, odmah proizlazi da je pripadno kritično područje određeno uvjetom

$$d(\widehat{t}) \geq H^{-1}_{r-v-1}(1-\alpha)\,. \tag{18}$$

U promatranom smo primjeru postavili hipotezu H_0 da podaci potječu od Poissonove razdiobe, u kojoj postoji jednodimenzionalni parametar $t = \lambda(\lambda > 0)$, tako da je $v = 1$ i stoga test-statistici $D(\widehat{\Lambda})$ iz (15) pripada hikvadrat-razdioba sa $r - v - 1 = 12 - 1 - 1 = 10$ stupnjeva slobode. Za $\alpha = 0{,}05$, kritično područje određeno je uvjetom $d(\widehat{\lambda}) \geq H^{-1}_{10}(0{,}95) = 18{,}3$ (v. tabl. VI. u Dodatku). Iz (14) se vidi da su podaci dali test-statistici vrijednost 9,57, što ne pada u kritično područje $[18{,}3\,;\infty\rangle$, a to znači da hipotezu H_0 treba prihvatiti.

3. Hipoteze o tipu vjerojatnosne razdiobe

U IX.2. je već prikazano kako se konstruira test za testiranje hipoteze o bilo kojoj diskretnoj razdiobi vjerojatnosti. Teorijska iznova za to nalazi se u navedenom Fisherovu teoremu. Odmah se postavlja pitanje da li se Fisherov teorem može primijeniti i pri konstrukciji testova o kontinuiranim razdiobama vjerojatnosti. Kako se to radi ilustrirat će se idućim primjerom.

3. primjer

Promatranjem statističkih podataka o tlačnoj čvrstoći betonskih kocki iz 5. primjera u I.4, te pripadnih tablica (tabl. 5. i 7. u I. pogl.) i grafikona (sl. 8, 9, 10, 11. u I. pogl.) može se pomisliti da izmjereni podaci možda potječu od neke normalne razdiobe $N(\mu, \sigma^2)$. Stoga se odmah postavlja pitanje kako konstruirati test za testiranje hipoteze H_0: (podaci potječu od normalne razdiobe), prema alternativnoj hipotezi H_1: (podaci ne potječu od normalne razdiobe). Budući da su podaci u tabl. 5. u I. poglavlju grupirani u 20 razreda, nameće se ideja da se za donošenje odluke iskoristi Fisherov teorem tako da se uzme $n = 100$, $r = 20$, za $\widehat{f}_j$ frekvencija j-tog razreda, a za $p_j(\widehat{\boldsymbol{t}})$ vjerojatnost j-tog razreda izračunana po normalnoj razdiobi $N(\widehat{\mu}, \widehat{\sigma}^2)$, gdje je $\widehat{\boldsymbol{t}} = (\widehat{\mu}, \widehat{\sigma}^2)$ vrijednost minimalnog hikvadratnog procjenitelja $\widehat{\boldsymbol{T}}$ za vektorski parametar $\boldsymbol{t} = (\mu, \sigma^2)$, koji se u ovom slučaju približno podudara s ML-procjeniteljem $\widehat{\boldsymbol{T}} = (\overline{X}, \widehat{\Sigma}^2)$ (v. VI.4).

Međutim, iz Pearsonova teorema je očigledno da se primjena hikvadrat-testa temelji na pretpostavci da je n dovoljno veliko da se, za svaki j, smije binomna razdioba $B(n, p_j)$ uzoračke frekvencije $\widehat{F}_j$ aproksimirati odgovarajućom normalnom razdiobom, a to je dopušteno samo za one p_j koji nisu preblizu nuli i jedinici, tj. ako očekivana (teorijska) frekvencija $np_j = f_j^{(0)}$ nije premalena, a ni preblizu n.

Smatra se da je za primjenu hikvadrat-testa nužno da svaki razred ima očekivanu (teorijsku) frekvenciju bar 5 ili više. Odmah se vidi da grupiranje podataka u razrede primijenjeno u tabl. 5. u I. poglavlju nije prikladno za primjenu hikvadrat-testa. Uzme li se, naime, $\widehat{\mu} = \overline{x} = 35$ i $\widehat{\sigma}^2 = 21$, što su približne vrijednosti ML-procjenitelja $\widehat{\boldsymbol{T}} = (\overline{X}, \widehat{\Sigma}^2)$ u promatranom primjeru, mogu se izračunati vjerojatnosti razreda u tabl. 5. u I.4. prema normalnoj razdiobi N(35,21). Koristeći se formulom (49) iz IV.5. dobiva se

$$p_j(\widehat{\boldsymbol{t}}) = p_j(\widehat{\mu}, \widehat{\sigma}^2) = \Phi\Big(\frac{a_j - \widehat{\mu}}{\widehat{\sigma}}\Big) - \Phi\Big(\frac{a_{j-1} - \widehat{\mu}}{\widehat{\sigma}}\Big)\,, \qquad j = 1, \ldots, 20\,,$$

gdje je a_j gornja granica, a a_{j-1} donja granica j-tog razreda. Pripadne teorijske frekvencije $f_j^{(0)}$, u konkretnom slučaju kada je $n = 200$, $\widehat{\mu} = 35$ i $\widehat{\sigma} = \sqrt{21}$, izražene su formulom

$$f_j^{(0)} = n\,p_j(\widehat{\boldsymbol{t}}) = 100 \cdot \Big[\Phi\Big(\frac{a_j - 35}{\sqrt{21}}\Big) - \Phi\Big(\frac{a_{j-1} - 35}{\sqrt{21}}\Big)\Big]\,. \tag{19}$$

Izvedu li se konkretni proračuni po formuli (19) s veličinama iz tabl. 5. u I.4, odmah se vidi da teorijske frekvencije za više razreda ne zadovoljavaju uvjet da su veće od 4. Tako je, na primjer, $f_1^{(0)} \approx 1{,}28$, $f_2^{(0)} \approx 2{,}13$ itd. Zato nije moguće primijeniti hikvadrat-test na statističke podatke 5. primjera u I.4, grupirane u tabl. 5.

No, isti se podaci mogu grupirati u razrede i na druge načine, o čemu je već bilo govora u I. poglavlju. Tako su u tabl. 7. u I.6. isti podaci grupirani u $r = 10$ razreda. Odmah se vidi da ni ta podjela na razrede ne zadovoljava uvjet da očekivana teorijska frekvencija svakog razreda bude 5 ili više. Stoga ćemo načiniti novu podjelu istih podataka na razrede, koja će zadovoljavati navedeni uvjet, a prikazana je u tabl. 2. Statistički podaci o tlačnoj čvrstoći betonskih kocki priređeni su u tabl. 2. tako da se može primijeniti hikvadrat-test za testiranje hipoteze H_0: (podaci potječu od normalne razdiobe), prema alternativnoj hipotezi

Tablica 2.

Redni broj razreda (j)	Razred		Teorijska frekvencija ($f_j^{(0)}$)	Empirijska frekvencija ($\widehat{f}_j$)	$\dfrac{(\widehat{f}_j - f_j^{(0)})^2}{f_j^{(0)}}$
	donja granica (a_{j-1})	gornja granica (a_j)			
1	$-\infty$	29,8	12,93	11	0,285
2	29,8	32,2	14,16	11	0,705
3	32,2	34,6	19,32	21	0,146
4	34,6	37,0	20,23	21	0,029
5	37,0	39,4	16,51	15	0,138
6	39,4	41,8	10,91	12	0,109
7	41,8	∞	6,94	9	0,611
					$d(\widehat{t}) = 2{,}024$

H_1: (podaci ne potječu od normalne razdiobe). Može se, naime, smatrati da su zadovoljeni uvjeti Fisherova teorema, pri čemu je $n = 100$, $r = 7$, dok je $\boldsymbol{t} = (\mu, \sigma^2)$ dvodimenzionalni vektorski parametar, tako da je $v = 2$, što sve zajedno povlači da test-statistika $D(\widehat{T}) \sim X^2(4)$.

Uzme li se $\alpha = 0{,}05$, kritično područje definirano je uvjetom $d(\widehat{\boldsymbol{t}}) \geq H_4^{-1}(0{,}95) = 9{,}49$, pa se odmah vidi da dobivena vrijednost 2,024 test-statistike ne pada u kritično područje $[9{,}49\,;\infty\rangle$, što znači da ne treba odbaciti hipotezu o normalnoj razdiobi slučajne varijable X (tlačna čvrstoća betonskih kocki).

Opći postupak za primjenu hikvadrat-testa na testiranje hipoteze H_0 da izmjereni podaci $x_1, \ldots, x_n$ potječu od zadane klase $\mathcal{P} = \{P_t : t \in \Theta\}$ vjerojatnosnih razdioba izgleda ovako:

1. Na temelju danih podataka određuje se minimalna hikvadratna procjena $\widehat{t}$ parametra t (obično se zamjenjuje ML-procjenom).
2. Brojevna os (skup $\mathbf{R}$) razbija se na disjunktne intervale (razrede) $I_1, \ldots, I_r$ ($r \in \mathbf{N}$) tako, da je očekivana (teorijska) frekvencija $f_j^{(0)} = n\,P_{\widehat{t}}(I_j) \geq 5$ ($j = 1, \ldots, r$).
3. Određuju se empirijske frekvencije $\widehat{f}_j$ (broj onih podataka danog niza koji padaju u razred I_j). Ako neki od podataka padne baš na granicu dvaju razreda, onda se u svakom od njih uzima 0,5 kao doprinos empirijskoj frekvenciji razreda.
4. Izračunava se vrijednost test-statistike
$$d(\widehat{t}) = \sum_{j=1}^{r} \frac{(\widehat{f}_j - f_j^{(0)})^2}{f_j^{(0)}}\,.$$
5. Izabire se razina značajnosti α i određuje uvjet za kritično područje
$$d(\widehat{t}) \geq H_{r-v-1}^{-1}(1-\alpha)\,,$$
gdje je v dimenzija parametra t.

Istaknimo odmah neka dobra i neka loša svojstva hikvadrat-testa pri odlučivanju o tipu vjerojatnosne razdiobe od koje potječu dani podaci.

Dobro mu je svojstvo, svakako, da se može primijeniti i na diskretne i na kontinuirane razdiobe vjerojatnosti, iako je strogo teorijski utemeljen samo za diskretne razdiobe.

Fisherovim teoremom utvrđen je utjecaj procjene nepoznatog parametra na razdiobu test-statistike, što nije poznato kod nekih drugih testova.

Primjena hikvadrat-testa na kontinuirane razdiobe vjerojatnosti zahtijeva grupiranje podataka u razrede (može se primijeniti i kod diskretnih), što je donekle proizvoljan postupak, koji nema strogo teorijsko opravdanje. U tom je slučaju problematičan i utjecaj procjene $\widehat{t}$ nepoznatog parametra t na razdiobu vjerojatnosti test-statistike $D(\widehat{T})$. Računa li se, naime, procjena $\widehat{t}$ bez primjene grupiranja podataka u razrede (v. (5) i (6) u II.1), preporučuje se (v. [4]) da se ne smanjuje broj stupnjeva slobode hikvadrat-razdiobe za dimenziju v (broj komponenata) vektorskog parametra t, već da se uzima $D(\widehat{T}) \sim X^2(r-1)$.

Što se tiče izbora broja razreda r, sa stajališta Fisherova teorema, bolje je imati malo razreda s velikim frekvencijama razreda, jer tada hikvadrat-razdioba dobro aproksimira razdiobu vjerojatnosti test-statistike $D(\widehat{T})$. Međutim, ako se velik broj podataka (n veliko) grupira u mali broj razreda (r maleno), onda se očigledno gubi znatan dio informacije sadržane u danom nizu podataka.

Kada bismo, na primjer, uzeli samo dva razreda $I_1 = \langle -\infty, 0]$ i $I_2 = \langle 0, \infty\rangle$ ($r = 2$) i postavili nul-hipotezu da podaci potječu od uniformne razdiobe $U\left(-\frac{t}{2}, \frac{t}{2}\right)$ ($t > 0$), imali bismo pripadne vjerojatnosti razreda $P_t(I_1) = P_t(I_2) = \frac{1}{2}$, odnosno pripadne teorijske frekvencije $f_1^{(0)} = f_2^{(0)} = \frac{n}{2}$. No iste bismo teorijske frekvencije dobili i kada bismo postavili nul-hipotezu da podaci potječu od normalne razdiobe $N(0, \sigma^2)$ ($\sigma > 0$). To znači da se u oba slučaja odluka donosi na temelju iste vrijednosti test-statistike i uz isto kritično područje, pa će se na temelju danih podataka i jedna i druga nul-hipoteza prihvaćati, odnosno odbacivati. Kaže se da tako konstruirani test ima slabu *razlučivost*, jer i s vrlo velikim brojem n podataka ne razlučuje uniformnu razdiobu od normalne.

4. Razlučivost hikvadrat-testa

Navede li nas hikvadrat-test da odbacimo hipotezu o pretpostavljenoj razdiobi vjerojatnosti, možemo biti prilično sigurni da izmjereni podaci ne potječu od pretpostavljene vjerojatnosne razdiobe. Navede li nas, međutim, da prihvatimo hipotezu, i ako još uzorak nije jako velik, onda ne smijemo smatrati da su druge razdiobe vjerojatnosti isključene.

Često se, naime, događa da se uz iste podatke, primjenom hikvadrat-testa, može dobiti zaključak o prihvaćanju više različitih vjerojatnosnih razdioba.

Stoga je prirodno da se postavi pitanje koliko se trebaju razlikovati dvije vjerojatnosne razdiobe da bi ih hikvadrat-test, uz danu veličinu uzorka n i razinu značajnosti α, mogao razlučiti. No time je odmah postavljeno i pitanje kako "mjeriti" razliku između vjerojatnosnih razdioba. Općenito je to vrlo složeno pitanje, međutim za posebni slučaj koji nas ovdje zanima zadatak se može riješiti nešto jednostavnije.

Ideja vodilja nazire se iz formule (4), gdje se veličina d može interpretirati kao mjera razlike (udaljenost, distanca) između teorijske razdiobe $\boldsymbol{p}^{(0)} = (p_1^{(0)}, \ldots, p_r^{(0)})$ i empirijske razdiobe $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \ldots, \widehat{p}_r)$. Zato se općenito čini razumnim veličinu

$$\Delta(\boldsymbol{p}^{(0)}, \boldsymbol{p}) = \sum_{j=1}^{r} \frac{(p_j - p_j^{(0)})^2}{p_j^{(0)}} \qquad (20)$$

nazvati razdaljinom diskretne vjerojatnosne razdiobe $\boldsymbol{p}^{(0)}$, koja zadovoljava uvjet $p_j^{(0)} > 0$ ($j = 1, \ldots, r$) od bilo koje diskretne vjerojatnosne razdiobe $\boldsymbol{p} = (p_1, \ldots, p_r)$ ($p_j \geq 0$, $\sum_{j=1}^{r} p_j = 1$).

Primijetimo da u (20) nije ispunjen uvjet simetrije $\Delta(p^{(0)}, p) = \Delta(p, p^{(0)})$, pa se ne može govoriti o međusobnoj udaljenosti vjerojatnosnih razdioba p i $p^{(0)}$.

Primijeni li se hikvadrat-test na testiranje hipoteze da dani niz od n podataka potječe od vjerojatnosne razdiobe $p^{(0)}$, uz razinu značajnosti α, može se promatrati skup $\mathcal{R}(\alpha, n)$ svih onih vjerojatnosnih razdioba p koje će se tim testom odbaciti. Veličina

$$\Delta_0 = \min_{p \in \mathcal{R}(\alpha,n)} \Delta(p^{(0)}, p) \tag{21}$$

zove se *minimalna razlučiva razdaljina* razine značajnosti α. Ona ima značenje minimalne razdaljine neke razdiobe p, od radiobe $p^{(0)}$, koja se hikvadrat-testom razine značajnosti α i uz veličinu uzorka n može razlučiti od razdiobe $p^{(0)}$.

Postoji (v. zad. 10) jednostavna približna formula koja vrijedi za dovoljno velike n i koja glasi

$$\Delta_0 = \frac{\gamma_r \sqrt{r-1}}{n}, \tag{22}$$

gdje je γ_r praktički neovisno o r i ovisi samo o α, tako da je za $5 \le r \le 400$ i $\alpha = 0{,}1$, $\gamma_r \approx 6$, a za $\alpha = 0{,}05$, $\gamma_r \approx 8$. Iz (22) se vidi da Δ_0 ne ovisi o polaznoj vjerojatnosnoj razdiobi $p^{(0)}$.

Ako je, na primjer, $r = 10$, $n = 100$ i $\alpha = 0{,}05$, onda se dobiva $\Delta_0 = 0{,}24$, što znači da se sve one razdiobe p, koje su od polazne razdiobe $p^{(0)}$ udaljene za manje od 0,24, prihvaćaju primjenom hikvadrat-testa s danim n i α, jednako kao i razdioba $p^{(0)}$. Drugim riječima, ako je stvarna razdioba od koje potječu podaci $p \neq p^{(0)}$, ali je $\Delta(p^{(0)}, p) < \Delta_0$, onda će se nul-hipoteza ipak prihvaćati s vjerojatnošću $1 - \alpha$, odnosno odbacivati s vjerojatnošću α, dakle isto kao da je stvarna razdioba $p^{(0)}$.

Može se reći da veličina Δ_0 karakterizira, na određeni način, moć razlučivanja hikvadrat-testa. Ako je Δ_0 maleno, razumno je reći da je moć razlučivanja velika, pa se stoga veličina

$$\delta_0 = \frac{1}{\Delta_0} = \frac{n}{\gamma_r \sqrt{r-1}} \tag{23}$$

zove *razlučivost hikvadrat-testa*.

Odmah se vidi da se razlučivost povećava s povećanjem veličine uzorka n, što je intuitivno vrlo prihvatljivo. Međutim, formula (23) pokazuje da se razlučivost smanjuje pri povećanju broja r razreda u koje su grupirani zadani podaci.

Opisani postupak može se primijeniti i za određivanje razlučivosti hikvadrat--testa kad se on primjenjuje na kontinuirane razdiobe vjerojatnosti. To će se ilustrirati idućim primjerom.

4. primjer

Da bi se odredila razdaljina između standardne normalne razdiobe $N(0, 1)$ i uniformne razdiobe s očekivanjem 0 i varijancom 1, tj. uniformne razdiobe $U(-\sqrt{3}, \sqrt{3})$, brojevna os, odnosno skup $\mathbf{R}$, razbit će se na $r = 10$ intervala (razreda) i zatim izračunati vjerojatnost svakog razreda po standardnoj normalnoj i uniformnoj razdiobi $U(-\sqrt{3}, \sqrt{3})$, što je prikazano u tabl. 3.

Primjenom formule (20) nalazi se da je $\Delta(p^{(0)}, p) = 0{,}1542$, što se može interpretirati tako da se kaže da je razdaljina uniformne razdiobe $U(-\sqrt{3}, \sqrt{3})$ od normalne razdiobe $N(0, 1)$ na razini danih $r = 10$ razreda, jednaka 0,1542.

Treba li naći minimalnu veličinu uzorka n, uz koju se s razinom značajnosti $\alpha = 0{,}05$ može razlučiti normalna razdioba $N(0, 1)$ od uniformne razdiobe $U(-\sqrt{3}, \sqrt{3})$, najprije će se, na temelju (22), zaključiti da je u tom slučaju minimalna razlučiva razdaljina $\Delta_0 = \frac{24}{n}$, te stoga treba n odrediti tako da bude zadovoljen uvjet

$$0{,}1542 > \frac{24}{n} \Rightarrow n > 155\,.$$

To nam pokazuje da se i s vrlo velikim uzorkom (recimo $n = 150$), primjenom hikvadrat-testa uz razinu značajnosti od 5 %, prihvaća normalna razdioba, iako je možda stvarna razdioba uniformna.

Tablica 3.

Redni broj razreda	Razred		$N(0,1)$ $p_j^{(0)}$	$U(-\sqrt{3},\sqrt{3})$ p_j	$\frac{(p_j^{(0)} - p_j)^2}{p_j^{(0)}}$
	donja granica	gornja granica			
1	$-\infty$	-2	0,023	0	0,0228
2	-2	$-1{,}5$	0,044	0,068	0,0131
3	$-1{,}5$	-1	0,092	0,144	0,0294
4	-1	$-0{,}5$	0,150	0,144	0,0002
5	$-0{,}5$	0	0,191	0,144	0,0116
6	0	0,5	0,191	0,144	0,0116
7	0,5	1	0,150	0,144	0,0002
8	1	1,5	0,092	0,144	0,0294
9	1,5	2	0,044	0,068	0,0131
10	2	∞	0,023	0	0,0228

Još se teže razlučuje normalna razdioba od Laplaceove razdiobe (v. točku 6. u V.6). Kada bi se izvelo slično računanje za $N(0,1)$ i Laplaceovu razdiobu parametra $\sqrt{2}$, koja ima očekivanje 0 i varijancu 1, našlo bi se da je nužno imati uzorak veličine $n > 450$ da bi se razlučile te dvije vjerojatnosne razdiobe uz razinu značajnosti $\alpha = 0{,}05$.

Za razlučivanje $N(0,1)$ i Cauchyjeve razdiobe (v. točku 7. u V.6), na istoj razini značajnosti $\alpha = 0{,}05$, nužno je raspolagati uzorkom veličine $n > 133$.

Na prvi pogled može izgledati neobično da se teže razlučuje normalna razdioba od uniformne nego od Cauchyjeve razdiobe, čija je krivulja razdiobe mnogo sličnija normalnoj krivulji nego krivulja uniformne razdiobe (usp. sl. 9. i 12 u IV.3). Uzrok tome je u činjenici da se normalna i Cauchyjeva razdioba jako razlikuju u "repnim" dijelovima (razredima kojima pripada mala vjerojatnost), a što mnogo pridonosi vrijednosti $\Delta(p^{(0)}, p)$, što se razabire iz (20). Općenita je konstantacija da je hikvadrat-test vrlo osjetljiv na "repno" ponašanje razdioba.

Na temelju provedenih razmatranja može se zaključiti da, primjenom hikvadrat--testa na uzorcima veličine $n < 100$, hipoteza o normalnoj razdiobi neće biti odbačena i ako se stvarna razdioba od koje potječu uzorački podaci razlikuje od

normalne razdiobe. Čak se i s uzorkom veličine $n = 400$ ne može razlučiti normalna razdioba od slične joj simetrične razdiobe, poput Laplaceove razdiobe, čiji se repovi također eksponencijalno približuju nuli.

Sve to upozorava da je potreban znatan oprez pri donošenju zaključaka o razdiobi vjerojatnosti uz primjenu hikvadrat-testa, posebno u uvjetima relativno male veličine uzorka.

5. Hipoteza o nezavisnosti

U vezi s dvodimenzionalnim razdiobama vjerojatnosti definiran je pojam stohastičke nezavisnosti slučajnih varijabli X i Y (v. V.1), pa se prirodno nameće zadatak da se konstruira test za testiranje hipoteze H_0 da su X i Y nezavisne slučajne varijable. Polazi se od toga da se, na temelju niza od n mjerenja (uređenih parova) $(x_1, y_1), \ldots, (x_n, y_n)$ slučajnog vektora (X, Y), može oblikovati pripadna kontingencijska tablica (v. III.1. i III.7), gdje je za vrijednosti slučajne varijable X predviđeno r, a za vrijednosti slučajne varijable Y predviđeno je s razreda. Ako je x_i palo u j-ti razred varijable X, a y_i u k-ti razred varijable Y, onda se kaže da je uređeni par (x_i, y_i) $(i = 1, \ldots, n)$ pao u polje (j, k) kontingencijske tablice $(j = 1, \ldots, r,\ k = 1, \ldots, s)$. Sa $\widehat{f}_{jk}$ označit će se frekvencija, a sa $\widehat{p}_{jk} = \frac{1}{n}\widehat{f}_{jk}$ relativna frekvencija polja (j, k) u danom nizu podataka $(x_1, y_1), \ldots, (x_n, y_n)$.

Očigledno vrijedi

$$\sum_{j=1}^{r}\sum_{k=1}^{s}\widehat{f}_{jk} = n\,, \qquad \sum_{j=1}^{r}\sum_{k=1}^{s}\widehat{p}_{jk} = 1\,. \tag{24}$$

Ako p_{jk} $(p_{jk} \geq 0,\ \sum_{j=1}^{r}\sum_{k=1}^{s} p_{jk} = 1)$ označuje teorijsku vjerojatnost koja pripada polju (j, k) kontingencijske tablice na temelju vjerojatnosne razdiobe slučajnog vektora (X, Y), tada se može govoriti i o teorijskoj (očekivanoj) frekvenciji polja (j, k) pri n-strukom ponavljanju nezavisnih mjerenja slučajnog vektora (X, Y)

$$f_{jk} = np_{jk}\,, \qquad \sum_{j=1}^{r}\sum_{k=1}^{s} f_{jk} = n\,. \tag{25}$$

Statistika $\widehat{F}_{jk}$, čije su vrijednosti $\widehat{f}_{jk}$, nepristrani je procjenitelj za nepoznati parametar f_{jk}. Ako je n dovoljno veliko, tako da su i f_{jk} $(j = 1, \ldots, r\ k = 1, \ldots, s)$ dovoljno veliki za primjenu svojstva asimptotske normalnosti procjenitelja $\widehat{F}_{jk}$, može se primijeniti Pearsonov teorem (v. (7)), tj. može se zaključiti da slučajna varijabla

$$D = \sum_{j=1}^{r}\sum_{k=1}^{s}\frac{(\widehat{F}_{jk} - f_{jk})^2}{f_{jk}} \sim X^2(rs - 1)\,. \tag{26}$$

To pokazuje da se hikvadrat-test može primijeniti i na dvodimenzionalni slučaj

testiranja hipoteze $H_0 : p_{jk} = p_{jk}^{(0)}$, prema alternativnoj hipotezi $H_1 : p_{jk} \neq p_{jk}^{(0)}$ za bar jedan uređeni par (j, k) $(j = 1, \ldots, r\ , k = 1, \ldots, s)$. Očigledna je i mogućnost generalizacije hikvadrat-testa na višedimenzionalni slučaj, u što se nećemo upuštati.

Pretpostavi li se da vjerojatnosti p_{jk}, a zbog (25) i teorijske frekvencije f_{jk}, ovise o nepoznatom parametru $\boldsymbol{t}(\boldsymbol{t} \in \boldsymbol{\Theta})$ pripadne dvodimenzionalne vjerojatnosne razdiobe $P_{\boldsymbol{t}}$, slučajna će varijabla

$$D(\boldsymbol{t}) = \sum_{j=1}^{r}\sum_{k=1}^{s}\frac{[\widehat{F}_{jk} - f_{jk}(\boldsymbol{t})]^2}{f_{jk}(\boldsymbol{t})} \tag{27}$$

upućivati na globalnu razliku između empirijskih i teorijskih frekvencija. Budući da je $\boldsymbol{t}$ nepoznato, u (27) će se umjesto $\boldsymbol{t}$ staviti pripadni minimalni hikvadratni procjenitelj $\widehat{\boldsymbol{T}}$ i tada, prema Fisherovu teoremu, vrijedi

$$D(\widehat{\boldsymbol{T}}) \sim X^2(rs - v - 1)\,, \tag{28}$$

gdje je kao i obično, v dimenzija vektorskog parametra $\boldsymbol{t}$. U tome je teorijska osnova za primjenu hikvadrat-testa pri testiranju hipoteza o tipu višedimenzionalne razdiobe vjerojatnosti.

Vratimo se, međutim, problemu testiranja hipoteze o nezavisnosti slučajnih varijabli X i Y. Ako je hipoteza o nezavisnosti istinita, onda vrijedi (v. (10) u V.2)

$$p_{jk} = p_j\, q_k\,, \qquad j = 1, \ldots, r\ , k = 1, \ldots, s\,, \tag{29}$$

gdje je

$$p_j = \sum_{k=1}^{s} p_{jk}\,, \quad q_k = \sum_{j=1}^{r} p_{jk}\,, \quad \sum_{j=1}^{r} p_j = \sum_{k=1}^{s} q_k = 1\,. \tag{30}$$

Definirajmo veličine $f_j = np_j$ i $g_k = nq_k$, pa se odmah vidi da se f_j može interpretirati kao teorijska frekvencija j-tog razreda prve varijable (X), a g_k kao teorijska frekvencija k-tog razreda druge varijable (Y).

Stavi li se

$$\widehat{f}_j = \sum_{k=1}^{s}\widehat{f}_{jk}\,, \qquad \widehat{g}_k = \sum_{j=1}^{r}\widehat{f}_{jk}\,, \tag{31}$$

vidi se da je $\widehat{f}_j$ empirijska frekvencija j-tog razreda iksova, a $\widehat{g}_k$ empirijska frekvencija k-tog razreda ipsilona, na danom nizu empirijskih podataka. Očigledno vrijedi

$$\sum_{j=1}^{r}\widehat{f}_j = \sum_{k=1}^{s}\widehat{g}_k = n\,. \tag{32}$$

Sada se može reći da se zadatak sastoji u testiranju hipoteze

$$H_0 : (p_{jk} = p_j\, q_k\,, \quad j = 1, \ldots, r\,, \quad k = 1, \ldots, s)\,,$$

prema alternativnoj hipotezi H_1 : (postoji bar jedan uređeni par (j, k) za koji je $p_{jk} \neq p_j\, q_k$).

Kada bi brojevi p_j i q_k, koji zadovoljavaju ograničenja (30), bili poznati, onda bi se zadatak mogao riješiti pomoću test-statistike (26), gdje bi se uzelo

$$f_{jk} = np_j q_k .$$

Međutim, ti su brojevi redovito nepoznati, tako da se moraju razmatrati kao nepoznati parametri, odnosno kao nepoznate komponente vektorskog parametra $\boldsymbol{t} = (p_1, \ldots, p_r, q_1, \ldots, q_s)$ dimenzije $r + s$.

Uzme li se kao vektorski procjenitelj za nepoznati parametar $\boldsymbol{t}$

$$\widehat{\boldsymbol{T}} = (\widehat{P}_1, \ldots, \widehat{P}_r, \widehat{Q}_1, \ldots, \widehat{Q}_s),$$

gdje je

$$\widehat{P}_j = \frac{1}{n} \widehat{F}_j , \qquad \widehat{Q}_k = \frac{1}{n} \widehat{G}_k ,$$

pri čemu su $\widehat{F}_j$ i $\widehat{G}_k$ statistike (uzoračke frekvencije) s vrijednostima $\widehat{f}_j$ i $\widehat{g}_k$, definiranim u (31), primjenom Fisherova teorema (v. (17)) zaključuje se da vrijedi

$$(33) \quad D(\widehat{\boldsymbol{T}}) = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{[\widehat{F}_{jk} - f_{jk}(\widehat{\boldsymbol{T}})]^2}{f_{jk}(\widehat{\boldsymbol{T}})} = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(\widehat{F}_{jk} - n\widehat{P}_j \widehat{Q}_k)^2}{n\widehat{P}_j \widehat{Q}_k} \sim \chi^2[(r-1)(s-1)] .$$

Očigledno je, naime, da komponente vektorskog parametra $\boldsymbol{t}$ nisu nezavisne, jer između njih postoje dvije funkcijske veze (v. (30)), tako da je $\boldsymbol{t}$, zapravo, nepoznati vektorski parametar dimenzije $v = r - 1 + s - 1 = r + s - 2$, što prema Fisherovu teoremu ima za posljedicu da test-statistika (33) ima hikvadrat-razdiobu sa $rs - v - 1 = (r-1)(s-1)$ stupnjeva slobode.

Iz (33) se razabire da se vrijednost test-statistike $D(\widehat{\boldsymbol{T}})$

$$(34) \qquad d = \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(\widehat{f}_{jk} - n\widehat{p}_j \widehat{q}_k)^2}{n\widehat{p}_j \widehat{q}_k} = n \left(\sum_{j=1}^{r} \sum_{k=1}^{s} \frac{\widehat{f}_{jk}^2}{\widehat{f}_j \widehat{g}_k} - 1 \right)$$

može interpretirati kao pokazatelj "udaljenosti" od nezavisnosti slučajnih varijabli X i Y, dobiven na danom n-članu nizu mjerenja slučajnog vektora (X, Y) i uz provedeno grupiranje podataka u razrede. Sada možemo odrediti kritično područje testa, uz zadanu razinu značajnosti α, uvjetom

$$(35) \qquad d \geq H_m^{-1}(1 - \alpha) , \qquad m = (r-1)(s-1) .$$

Prema tome, hipoteza o nezavisnosti odbacuje se onda, kada se na danim podacima dobije "udaljenost" d veća od kritične vrijednosti $H_m^{-1}(1-\alpha)$. Za praktičnu primjenu promatranoga testa važno je još primijetiti da su uvjeti Fisherova teorema praktički zadovoljeni ako svakom polju kontingencijske tablice pripada frekvencija koja nije manja od 10. Iz formule (34) se, nadalje, vidi da vrijednost test-statistike ovisi samo o frekvencijama, a ne i o vrijednostima slučajnih varijabli X i Y, tako da se opisani test može primijeniti i za testiranje hipoteza o nezavisnosti nenumeričkih statističkih obilježja.

5. primjer

U 2. primjeru u III.6. navedena je kontingencijska tablica (tabl. 5), gdje je $r = 2$ i $s = 3$ i u kojoj su dane empirijske frekvencije odgovarajućih polja te tablice, a koje se odnose na istraživanje veze između povišenoga krvnog tlaka i pušenja. Krvni tlak (X) tretiran je kao nenumeričko obilježje klasificirano u $r = 2$ razreda (normalni i povišeni), a pušenje (Y) kao nenumeričko obilježje klasificirano u $s = 3$ razreda (nepušač, blagi pušač, teški pušač).

Usporedi li se veličina f^2 (formula (39) u III.6) s veličinom d iz (34), odmah se vidi da je $d = nf^2$. Budući da je u 2. primjeru iz III.6. $n = 180$ i $f^2 = 0{,}08$, onda je odgovarajuća vrijednost test-statistike $d = 180 \cdot 0{,}08 = 14{,}4$. Uzme li se $\alpha = 0{,}01$, nalazi se da je $m = 2$ i $H_2^{-1}(0{,}99) = 9{,}21$ (v. tabl. VI. u Dodatku), pa se vidi da dobivena vrijednost test-statistike pada u kritično područje $[9{,}21 ; \infty\rangle$, što znači da hipotezu H_0 o nezavisnosti promatranih veličina X i Y treba odbaciti. U konkretnom primjeru to znači da izmjereni podaci upućuju na postojanje određene statističke zavisnosti između krvnog tlaka i pušenja.

6. Hipoteza o jednakosti vjerojatnosnih razdioba

Mnogi praktični problemi svode se na zadatak testiranja hipoteze o jednakosti dvije vjerojatnosne razdiobe ili više njih. Riječ je, zapravo, o tome da se na temelju dvaju nizova statističkih podataka $x_1, \ldots, x_m$ i $y_1, \ldots, y_n$, donese odluka o tome da li oni potječu od iste teorijske razdiobe vjerojatnosti, ili od različitih. Praktički gledano, problem je donekle sličan problemima opisanim u 8. i 9. primjeru u VIII.7, gdje se polazilo od pretpostavke da je riječ o podacima koji potječu od normalne razdiobe, pa je trebalo testirati hipotezu o eventualnoj jednakosti odgovarajućih parametara (μ i σ^2). Sada se neće isticati pretpostavka o određenoj parametarskoj klasi vjerojatnosnih razdioba, već će se pokazati da se, uz vrlo općenite pretpostavke, problem može riješiti primjenom hikvadrat-testa.

Pretpostavimo da je riječ o diskretnoj razdiobi vjerojatnosti s konačnim skupom vrijednosti $A = \{a_1, \ldots, a_r\}$ i pripadnim vjerojatnostima $p_j \geq 0$, $\sum\limits_{j=1}^{r} p_j = 1$ (v. IV.1. i IV. 2). Ako su u nizu podataka $x_1, \ldots, x_m$ dobivene frekvencije $\widehat{f}_j$, a u nizu $y_1, \ldots, y_n$ frekvencije $\widehat{g}_j$ vrijednosti $a_j \in A$, onda vrijedi

$$(36) \qquad \sum_{j=1}^{r} \widehat{f}_j = m , \qquad \sum_{j=1}^{r} \widehat{g}_j = n .$$

Za odgovarajuće relativne frekvencije $\widehat{p}_j = \frac{1}{m} \widehat{f}_j$ i $\widehat{q}_j = \frac{1}{n} \widehat{g}_j$, dakako, vrijedi

$$(37) \qquad \sum_{j=1}^{r} \widehat{p}_j = \sum_{j=1}^{r} \widehat{q}_j = 1 .$$

Da bi se konstruirao test za testiranje hipoteze H_0: (oba niza podataka potječu

od iste vjerojatnosne razdiobe), prema alternativnoj hipotezi H_1: (nizovi podataka ne potječu od iste vjerojatnosne razdiobe), treba definirati prikladnu test-statistiku, koja će omogućiti definiranje kritičnog područja testa, uz zadanu razinu značajnosti α. Jedna od mogućnosti da se to postigne izgleda ovako: Pretpostavi se da prvi niz podataka $x_1, \ldots, x_m$ potječe od određene diskretne razdiobe vjerojatnosti, pa se na temelju Pearsonova teorema zaključuje da, za velike m, vrijedi

$$D_1 = \sum_{j=1}^{r} \frac{(\widehat{P}_j - p_j)^2}{p_j} m \sim \chi^2(r-1)\,. \tag{38}$$

Također se pretpostavi i da drugi niz podataka $y_1, \ldots, y_n$ potječe od iste vjerojatnosne razdiobe, zbog čega , za velike n, vrijedi

$$D_2 = \sum_{j=1}^{r} \frac{(\widehat{Q}_j - p_j)^2}{p_j} n \sim \chi^2(r-1)\,. \tag{39}$$

Slučajne varijable D_1 i D_2 su nezavisne, pa zato vrijedi (v. točku 4. u V.6)

$$D_0 = D_1 + D_2 \sim \chi^2(2r-2)\,. \tag{40}$$

Statistika D_0 čini se vrlo prikladnom za indiciranje valjanosti hipoteze H_0, međutim nezgoda je u tome što nisu poznati parametri $p_1, \ldots, p_r$ koji se, dakako, mogu razmatrati kao komponente vektorskog parametra $\boldsymbol{t} = (p_1, \ldots, p_r)$. Čini se, stoga, razložnim zamijeniti te parametre u (40) njihovim procjeniteljima, ali na uzorku veličine $m + n$, koji je dobiven spajanjem danih dvaju nizova podataka. Prema tome p_j će se zamijeniti sa

$$\overline{P}_j = \frac{\widehat{F}_j + \widehat{G}_j}{m+n}\,, \tag{41}$$

gdje su $\widehat{F}_j$ i $\widehat{G}_j$ statistike s vrijednostima $\widehat{f}_j$ i $\widehat{g}_j$. Tada će odgovarajuća test-statistika izgledati

$$D = \sum_{j=1}^{r} \left[\frac{(\widehat{P}_j - \overline{P}_j)^2}{\overline{P}_j} m + \frac{(\widehat{Q}_j - \overline{P}_j)^2}{\overline{P}_j} n \right] = \frac{mn}{m+n} \sum_{j=1}^{r} \frac{(\widehat{P}_j - \widehat{Q}_j)^2}{\overline{P}_j}\,,$$

odnosno

$$D = \frac{1}{mn} \sum_{j=1}^{r} \frac{(n\widehat{F}_j - m\widehat{G}_j)^2}{\widehat{F}_j + \widehat{G}_j}\,. \tag{42}$$

Ako je hipoteza H_0 stvarno istinita onda, u skladu s Fisherovim teoremom, test-statistici D iz (42) pripada hikvadrat-razdioba sa $2r - 2 - (r-1) = r - 1$ stupnjeva slobode. To omogućuje da se kritično područje razine značajnosti α odredi uvjetom

$$d \geq H_{r-1}^{-1}(1-\alpha)\,, \tag{43}$$

gdje je d vrijednost test-statistike D iz (42).

6. primjer

Za obradu određenoga nastavnog gradiva primijenjene su dvije različite nastavne metode. Metoda M_1 primijenjena je u skupini A od 100 učenika, a metoda M_2 u skupini B od 200 učenika. Da bi se utvrdio učinak, svi su učenici ispitani i ocijenjeni odgovarajućom ocjenom od 1 do 5. Dobiveni rezultati prikazani su u tabl. 4. Može li se, na temelju rezultata iz tabl. 4, smatrati da su obje nastavne metode jednakog učinka?

Tablica 4.

	Ocjena					$\sum$
	1	2	3	4	5	
Skupina A	14	26	34	16	10	100
Skupina B	18	36	58	56	32	200
$\sum$	32	62	92	72	42	300

Da bi se na ovaj zadatak mogao primijeniti hikvadrat-test, pretpostavit će se da postoji određena diskretna razdioba vjerojatnosti koja opisuje statističku zakonitost razdiobe frekvencija pojedinih ocjena u određenoj populaciji učenika pri usvajanju određene nastavne građe. Tada se postavljeno pitanje može formulirati i kao problem testiranja hipoteze H_0:(nizovi frekvencija skupina A i B potječu od iste teorijske razdiobe), prema alternativnoj hipotezi H_1:(ne potječu od iste razdiobe).

U skladu s uvedenim oznakama imamo $m = 100$, $n = 200$ i $r = 5$, a odgovarajućim proračunom dobiva se vrijednost test-statistike $d = 9{,}88$. Uzme li se $\alpha = 0{,}05$, iz tabl. VI. u Dodatku očitava se $H_4^{-1}(0{,}95) = 9{,}49$, pa se iz (43) zaključuje da dobivena vrijednost test-statistike pada u kritično područje, što znači da hipotezu H_0 treba odbaciti. Pri praktičnoj interpretaciji toga rezultata reći ćemo da dobiveni podaci iz tabl. 4. upućuju na značajne razlike u pogledu učinka između nastavnih metoda M_1 i M_2.

Važno je stalno imati na umu da hikvadrat-test zahtijeva velike uzorke, što znači da se opisani test smije primjenjivati samo za velike m i velike n. Praktički to znači da se, u slučaju pojave frekvencija manjih od 5, treba izvršiti grupiranje podataka u razrede, čime se, naravno, smanjuje r. Grupiranje u razrede je, dakako, nužno ako je riječ o kontinuiranoj razdiobi vjerojatnosti.

Kada nas hikvadrat-test navede na zaključak da se odbaci nul-hipoteza o jednakosti razdioba, možemo biti prilično uvjereni da smo ispravno postupili. Kada nas, međutim, navede da prihvatimo hipotezu o jednakosti razdioba, a da pritom i nismo raspolagali s naročito velikim uzorcima, onda ne smijemo biti previše čvrsto uvjereni da smo ispravno postupili.

Slično kao u IX.4, može se postaviti pitanje o minimalnoj razdaljini Δ_0 između dviju vjerojatnosnih razdioba, koja omogućuje njihovo razlučivanje pomoću hikvadrat-testa uz zadanu razinu značajnosti α i zadanu veličinu uzorka. Ako su $\boldsymbol{p}^{(1)} = (p_1^{(1)}, \ldots, p_r^{(1)})$ i $\boldsymbol{p}^{(2)} = (p_1^{(2)}, \ldots, p_r^{(2)})$ dvije diskretne razdiobe vjerojatnosti i ako se stavi

$$p_j = \frac{1}{2}\,(p_j^{(1)} + p_j^{(2)})\,, \qquad \Delta p_j = p_j^{(1)} - p_j^{(2)}\,, \qquad j = 1, \ldots, r\,,$$

onda se veličina

$$\Delta(p^{(1)}, p^{(2)}) = \sum_{j=1}^{r} \frac{(\Delta p_j)^2}{p_j} \tag{44}$$

zove *razdaljina između* $p^{(1)}$ i $p^{(2)}$. Ako je $m = n$, tada za velike n približno vrijedi (v. [4])

$$\Delta_0 = \frac{2}{n}\,\gamma_r\sqrt{r-1}\,, \tag{45}$$

gdje za γ_r vrijedi sve ono što je rečeno u vezi s formulom (22).

Tako se, na primjer, pomoću hikvadrat-testa ne mogu razlučiti dvije diskretne razdiobe, za koje je $r = 10$, uz razinu značajnosti $\alpha = 0{,}05$, pomoću dva uzorka veličine $n = 100$, ako je njihova razdaljina manja od $\Delta_0 = 0{,}48$. Ne bi se, na primjer, mogla razlučiti uniformna diskretna razdioba

$$p_1^{(1)} = p_2^{(1)} = p_3^{(1)} = p_4^{(1)} = p_5^{(1)} = p_6^{(1)} = p_7^{(1)} = p_8^{(1)} = p_9^{(1)} = p_{10}^{(1)} = 0{,}1$$

od diskretne razdiobe

$$p_1^{(2)} = p_3^{(2)} = p_5^{(2)} = p_7^{(2)} = p_9^{(2)} = 0{,}08$$
$$p_2^{(2)} = p_4^{(2)} = p_6^{(2)} = p_8^{(2)} = p_{10}^{(2)} = 0{,}12\,.$$

Za njihovo razlučivanje nužno je imati uzorke veličine $n > 1\,200$.

7. Hipoteza o homogenosti

Hikvadrat-test može se upotrijebiti i za otkrivanje određenih nestabilnosti u nekom procesu koji se prati mjerenjem relevantne veličine X, za koju se pretpostavlja da u stabilnim uvjetima ima fiksiranu razdiobu vjerojatnosti.

Tako se, na primjer, proces generiranja slučajnih brojeva (generator slučajnih brojeva) može shvatiti kao nezavisno ponavljanje mjerenja (opažanja) diskretne slučajne varijable X, sa skupom vrijednosti $A = \{0,1,2,3,4,5,6,7,8,9\}$ i pripadnim vjerojatnostima $p_0 = p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8 = p_9 = 0{,}1$, pa se odmah postavlja pitanje provjere stabilnosti toga procesa. To se obično rješava tako da se uzme vrlo dugi niz, recimo $n = 2\,000$, generiranih slučajnih brojeva i zatim razbije taj niz na određeni broj, recimo $m = 10$, podnizova. Želimo li imati jednakobrojne podnizove, svaki će sadržavati 200 članova. Uspije li se dokazati da svih m podnizova potječu od iste diskretne razdiobe vjerojatnosti, smatrat će se da nema razloga sumnjati u stabilnost procesa generiranja slučajnih brojeva. Može se, zapravo, reći da promotreni podaci (njih n) kao cjelina podsjeduju određenu homogenost, koja se sastoji u podvrgavanju jednoj te istoj vjerojatnosnoj razdiobi, koja, dakako, i ne mora biti baš diskretna uniformna razdioba.

Odmah se nameće ideja da se problem utvrđivanja stabilnosti procesa, odnosno homogenosti niza danih statističkih podataka, postavi kao određeni problem testiranja statističkih hipoteza.

Općenito se može postaviti zadatak da se konstruira test za utvrđivanje homogenosti skupa od n podataka, kada se skup sastoji od m nizova, pri čemu prvi niz ima n_1, drugi n_2 itd. do m-tog niza koji ima n_m članova. To će se zapisati

$$\begin{matrix} x_{11},\ldots,x_{1n_1} \\ x_{21},\ldots,x_{2n_2} \\ \vdots \quad\quad \vdots \\ x_{m1},\ldots,x_{mn_m} \end{matrix} \tag{46}$$

pa je riječ o tome da se testira nul-hipoteza H_0: (svi nizovi potječu od iste diskretne vjerojatnosne razdiobe), prema alternativnoj hipotezi H_1: (ne potječu od iste razdiobe).

Očigledno je da se ovaj zadatak može shvatiti kao određena generalizacija problematike prethodnog poglavlja, gdje je, zapravo, bio posrijedi posebni slučaj $m = 2$. Zadržat ćemo isti sustav označivanja, pa će $\widehat{f}_j^{(i)}$ označivati frekvenciju vrijednosti a_j $(j = 1,\ldots,r)$ u i-tom nizu podataka $x_{i1},\ldots,x_{in_i}$ $\left(\sum_{j=1}^{r} \widehat{f}_j^{(i)} = n_i\,,\ \sum_{i=1}^{m} n_i = n\right)$, a $\widehat{p}_j^{(i)} = \frac{1}{n_i}\widehat{f}_j^{(i)}$ odgovarajuću relativnu frekvenciju $\left(\sum_{j=1}^{r} \widehat{p}_j^{(i)} = 1\right)$. Simbol $\widehat{P}_j^{(i)}$ označivat će procjenitelj za nepoznati parametar p_j, s vrijednostima $\widehat{p}_j^{(i)}$, pa će za velike n, prema Pearsonovu teoremu, vrijediti

$$D_i = \sum_{j=1}^{r} \frac{(\widehat{P}_j^{(i)} - p_j)^2}{p_j}\,n_i \sim \chi^2(r-1)\,, \qquad i = 1,\ldots,m\,, \tag{47}$$

i dalje

$$D_0 = \sum_{i=1}^{m} D_i \sim \chi^2[m(r-1)]\,. \tag{48}$$

Slučajna varijabla D_0 ne može poslužiti kao test-statistika za testiranje H_0, prema H_1, jer nisu poznati parametri $p_1,\ldots,p_r$. Zato će se p_j zamijeniti u (47) procjeniteljem $\overline{P}_j$, čije su vrijednosti

$$\overline{p}_j = \frac{1}{n}\sum_{i=1}^{m} \widehat{f}_j^{(i)}\,, \qquad j = 1,\ldots,r\,, \tag{49}$$

tj. relativne frekvencije vrijednosti a_j na uzorku veličine $n = n_1 + \ldots + n_m$ dobivenom spajanjem svih m nizova podataka u jedinstveni slučajni uzorak. Time se dobiva test-statistika

$$D = \sum_{i=1}^{m}\sum_{j=1}^{r} \frac{(\widehat{P}_j^{(i)} - \overline{P}_j)^2}{\overline{P}_j}\,n_i\,, \tag{50}$$

koja u uvjetima istinitosti hipoteze H_0, prema Fisherovu teoremu, ima hikvadrat-razdiobu sa $m(r-1) - (r-1) = (m-1)(r-1)$ stupnjeva slobode. Kritično

područje, razine značajnosti α, određeno je uvjetom

$$d \geq H_v^{-1}(1-\alpha)\,, \qquad v=(m-1)(r-1)\,, \tag{51}$$

gdje je d vrijednost test-statistike D iz (50).

7. primjer

U tvorničkom pogonu proizvode se televizori. Svakoga radnog dana u tjednu registrira se broj neispravnih televizora. Provedena su opažanja tijekom $n = 753$ dana i rezultati su prikazani u tabl. 5. U nutarnja polja tablice 5. upisane su frekvencije pojave j ($j = 0,1,2,3,4,5,6,7,8$ i više) neispravnih televizora dotičnog dana.

Tablica 5.

Broj (j) neispravnih televizora	Dan u tjednu (i)					$\sum$
	ponedjeljak (1)	utorak (2)	srijeda (3)	četvrtak (4)	petak (5)	
0	4	6	9	11	5	35
1	40	27	36	33	30	166
2	16	30	16	26	15	103
3	34	27	29	25	30	145
4	23	19	16	18	23	99
5	15	16	15	10	16	72
6	8	10	13	8	13	52
7	6	11	8	13	5	43
8 i više	6	5	7	10	10	38
$\sum$	152	151	149	154	147	753

Može li se, na temelju podataka iz tabl. 5, zaključiti da nema značajne razlike u pojavi neispravnih televizora tijekom tjedna, ili pak treba zaključiti da se proces proizvodnje značajno razlikuje u tom pogledu po danima u tjednu.

Odgovor na to pitanje može se dobiti primjenom upravo opisanog testa homogenosti. Usvaja se, dakle, matematički model u kojem se kao relevantna veličina X uzima dnevni broj neispravnih proizvoda, za koju je načinjeno $n = 753$ mjerenja, raspoređenih u $m = 5$ nizova (klasa), pri čemu prva klasa (ponedjeljak) sadrži $n_1 = 152$, druga klasa (utorak) $n_2 = 151$, treća klasa (srijeda) $n_3 = 149$, četvrta klasa (četvrtak) $n_4 = 154$ i peta klasa (petak) $n_5 = 147$ mjerenja (podataka). Testirat će se hipoteza H_0:(podaci u svih pet klasa potječu od iste vjerojatnosne razdiobe), prema alternativnoj hipotezi H_1:(ne potječu od iste razdiobe).

Da bi se izračunala vrijednost d test-statistike D iz (50), primijetimo da nutarnja polja tabl. 5. sadrže empirijske frekvencije $\widehat{f}_j^{(i)}$ ($i = 1,2,3,4,5$, $j = 0,1,2,3,4,5,6,7,8$), pa se lako izračunaju i konkretne vrijednosti veličina $\widehat{p}_j^{(i)}$ i $\overline{p}_j$, nužnih za proračun veličine $d = 32{,}7$.

Iz činjenice da je $r = 9$ i $m = 5$ proizlazi da test-statistici D pripada hikvadrat-razdioba sa $v = 32$ stupnja slobode. Uzme li se $\alpha = 0{,}05$, kritično područje

određeno je nejednakošću

$$d \geq H_{32}^{-1}(0{,}95) = 45{,}2\,,$$

iz čega se vidi da dobivena vrijednost $d = 32{,}7$ test-statistike ne pada u kritično područje, pa hipotezu H_0 treba prihvatiti.

Praktički bismo dobiveni rezultat interpretirali tako, da ne postoje značajne razlike u pojavi neispravnih televizora u različitim danima u tjednu.

Zadaci

1. Na temelju podataka iz zad.1. u I. poglavlju testirajte hipotezu da podaci potječu od idealne igraće kocke.

2. Može li se, na temelju podataka iz 1. primjera u I.1. (v. tabl. 2), zaključiti da u promatranoj učeničkoj populaciji ima po 10 % odličnih i slabih učenika, po 20 % vrlo dobrih i dovoljnih, te 40 % dobrih učenika, glede nastavnog predmeta matematike?

3. Može li se, na temelju podataka iz 4. primjera u I.3. (v. tabl. 4), zaključiti da u tekstovima hrvatskog jezika ima 50 % suglasnika?

4. Koristeći se podacima iz zad. 2. u I. poglavlju testirajte hipotezu da podaci potječu od binomne razdiobe $\mathrm{B}(r,p)$, gdje je $r = 20$ (ukupni broj učenika u razredu), dok je p nepoznati parametar (označuje vjerojatnost da bilo koji učenik izostane sa sata matematike). Ako je potrebno, provedite grupiranje podataka u razrede.

5. Može li se, na temelju podataka iz zad. 5. u I. poglavlju zaključiti da se dnevni broj prodanih pari cipela u promatranoj prodavaonici podvrgava Poissonovoj razdiobi?

6. Može li se, na temelju podataka iz zad. 8. u I. poglavlju zaključiti da se vlačna čvrstoća čelične žice podvrgava normalnoj razdiobi?

7. Može li se, na temelju podataka iz zad. 9. u I. poglavlju, zaključiti da je vijek trajanja promatranih žarulja slučajna varijabla:

a) eksponencijalne razdiobe,

b) lognormalne razdiobe?

8. Može li se, uz 1 % rizika odbacivanja istinite hipoteze, zaključiti da podaci potječu od normalne razdiobe, ako se uzme niz podataka iz:

a) zad. 11. u I. pogl. (težine novorođenčadi),

b) zad. 12. u I. pogl. (visine dvadesetogodišnjaka),

c) zad. 13. u I. pogl. (tlačne čvrstoće cementnih kocki)?

Kako izgleda zaključak ako se usvoji rizik od 10 % za odbacivanje istinite hipoteze?

9. Može li se, na temelju podataka iz zad. 14. u I. poglavlju, zaključiti da je vrijeme potrebno za popravak stroja slučajna varijabla eksponencijalne razdiobe?

10. Dokažite približnu formulu (22), polazeći od činjenice da je $\Delta(p^{(0)}, p) = nd$, gdje je d vrijednost test-statistike hikvadrat-testa.
Uputa: Upotrijebite tabl. VI. u Dodatku.

11. Primjenom formule (20), uz $r = 10$, nađite razdaljinu:

a) Laplaceove razdiobe parametra $\alpha = \sqrt{2}$ od N(0,1),
b) Cauchyjeve razdiobe od N(0,1).

12. Kolika je minimalna veličina uzorka n, uz koju se, s razinom značajnosti $\alpha = 0{,}05$, može razlučiti normalna razdioba s očekivanjem $\mu = \sqrt{e}$ i varijancom σ^2 =e(e-1) (e≈ 2,71828...) od lognormalne razdiobe s istim očekivanjem $\sqrt{e}$ i istom varijancom e(e-1)?

13. Može li se, na temelju podataka iz zad. 1. u III. poglavlju, zaključiti da su ishodi na prvoj (X) i drugoj (Y) igraćoj kocki nezavisne slučajne varijable?

14. Može li se, na temelju podataka iz zad. 5. u III. poglavlju, smatrati da su nenumerička statistička obilježja X i Y nezavisna?

15. Prikupite podatke kako je opisano u zad. 6. u III. poglavlju i ustanovite jesu li susjedna slova u tekstovima hrvatskog jezika stohastički zavisna.

16. Može li se, na temelju podataka iz zad. 1. u III. poglavlju, zaključiti da prvi niz podataka (x) i drugi niz podataka (y) potječu od iste diskretne razdiobe vjerojatnosti?

X. Prilagodba teorijske razdiobe empirijskim podacima

1. Empirijska funkcija razdiobe

Jedan od važnijih problema teorije statističkog zaključivanja svakako je problem procjene f.r.v. $F(x) = \mathrm{P}(X \leq x)$, $x \in \mathbf{R}$, promatrane slučajne varijable X. Ako pretpostavimo da je $x \in \mathbf{R}$ fiksirani broj, onda se $F(x)$ može shvatiti kao nepoznati parametar sa značenjem vjerojatnosti događaja da slučajna varijabla X poprimi vrijednost koja nije veća od broja x. Problem procjene vjerojatnosti događaja razmotren je u VII.5, gdje je utvrđeno da je relativna frekvencija uočenog događaja u danom nizu podataka $x_1, \ldots, x_n$ (mjerenja slučajne varijable X) nepristran, konzistentan i asimptotski normalan procjenitelj za nepoznatu vjerojatnost događaja.

Transformira li se niz podataka $x_1, \ldots, x_n$ u niz nula i jedinica $y_1, \ldots, y_n$ tako da se stavi

$$y_i = \begin{cases} 0, & \text{za} \quad x_i > x \\ 1, & \text{za} \quad x_i \leq x, \end{cases}$$

očigledno je veličina

$$F_n(x) = \frac{1}{n}(y_1 + \ldots + y_n) \tag{1}$$

relativna frekvencija događaja $\{X \leq x\}$ u danom nizu podataka i ona se može shvatiti kao vrijednost statistike

$$\widehat{F}_n(x) = \frac{1}{n}(Y_1 + \ldots + Y_n), \tag{2}$$

za koju se može reći da je nepristran, konzistentan i asimptotski normalan procjenitelj za nepoznatu vrijednost $F(x)$ f.r.v. u točki $x \in \mathbf{R}$.

Sada se prirodno nameće ideja da se promatra funkcija

$$x \mapsto F_n(x), x \in \mathbf{R}, \tag{3}$$

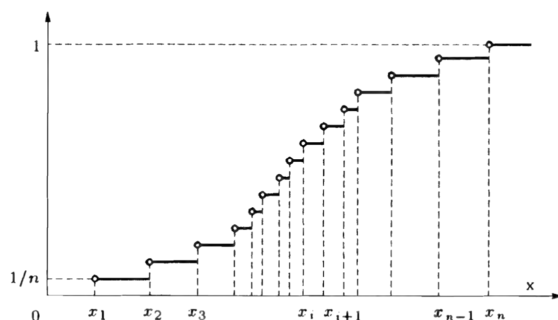
koja se zove ***empirijska*** ili ***uzoračka funkcija razdiobe*** za dani niz podataka, te da se istraži njezin odnos prema f.r.v. $x \mapsto F(x)$, $x \in \mathbf{R}$, slučajne varijable X.

Uočimo najprije da je empirijska funkcija razdiobe, zapravo, f.r.v. određene diskretne razdiobe vjerojatnosti, koja u svakoj točki $x_i \in \mathbf{R}$ ima "skok" visine $\frac{1}{n}$. Ako je u točku x_i palo više, recimo k podataka, onda taj skok iznosi $\frac{k}{n}$.

Pretpostavi li se da su podaci poredani po veličini, tj. da vrijedi $x_1 < x_2 < \ldots < x_n$, može se pisati

$$(4)\qquad F_n(x) = \begin{cases} 0 & , \text{ za } -\infty < x < x_1 \\ \frac{1}{n} & , \text{ za } x_1 \le x < x_2 \\ \frac{2}{n} & , \text{ za } x_2 \le x < x_3 \\ \vdots & \quad \vdots \\ \frac{i}{n} & , \text{ za } x_i \le x < x_{i+1} \\ \vdots & \quad \vdots \\ \frac{n-1}{n} & , \text{ za } x_{n-1} \le x < x_n \\ 1 & , \text{ za } x_n \le x < \infty. \end{cases}$$

Iz (4) se vidi da je graf funkcije F_n stepenasta krivulja sa skokovima veličine $\frac{1}{n}$ (v. sl. 27).



Slika 27. Tipični graf empirijske funkcije razdiobe

Pretpostavi li se još da je n veliko, može se primijeniti asimptotska normalnost procjenitelja $\widehat{F}_n(x)$ i rezultati izvedeni u VII.5, koji govore da se s vjerojatnošću γ $(0 < \gamma < 1)$ može jamčiti da će interval $\langle G_1, G_2 \rangle$ sa slučajnim granicama

$$(5)\qquad \begin{cases} G_1 = \widehat{F}_n(x) - z_\gamma \sqrt{\frac{1}{n}\widehat{F}_n(x)[1-\widehat{F}_n(x)]} \\ G_2 = \widehat{F}_n(x) + z_\gamma \sqrt{\frac{1}{n}\widehat{F}_n(x)[1-\widehat{F}_n(x)]} \end{cases}$$

pokriti nepoznatu vrijednost $F(x)$ f.r.v. slučajne varijable X. Iz (5) se razabire da širina toga intervala nije veća od $\frac{1}{\sqrt{n}}z_\gamma$, što pokazuje da se ona može, izborom veličine uzorka n, učiniti po volji malenom.

Odmah primijetimo da veličina $\frac{1}{\sqrt{n}}z_\gamma$ ne ovisi o $x \in \mathbf{R}$, pa se može reći da vrijednost $F_n(x)$ empirijske funkcije razdiobe svagdje (za svaki $x \in \mathbf{R}$) dobro aproksimira nepoznatu vrijednost $F(x)$ f.r.v. slučajne varijable X. U tom se smislu slučajna funkcija $x \mapsto \widehat{F}_n(x), x \in \mathbf{R}$, može smatrati procjeniteljem za nepoznatu f.r.v. $x \mapsto F(x), x \in \mathbf{R}$. Tome u prilog ide i poznati *Glivenko-Cantellijev teorem* (v.[26]) izražen relacijom[1]

$$(6)\qquad \mathrm{P}(\max_{x\in\mathbf{R}} |\widehat{F}_n(x) - F(x)| \to 0) = 1.$$

Relacija (6) može se protumačiti tako da, za velike n $(n \to \infty)$, maksimalna razlika između empirijske i teorijske funkcije razdiobe teži k nuli s vjerojatnošću jedan, tj. za gotovo sve moguće n-člane nizove podataka $x_1, \dots, x_n$ dobivene mjerenjem slučajne varijable X.

Relacija (6) jamči da za velike n empirijska funkcija razdiobe $x \mapsto F_n(x)$ postaje ne samo lokalno (za fiksirani $x \in \mathbf{R}$), već i globalno (za gotovo sve $x \in \mathbf{R}$) "bliska" nepoznatoj f.r.v. $x \mapsto F(x)$ i u tom smislu se funkcija F_n shvaća kao konkretna procjena za nepoznatu funkciju F, dobivena na temelju danog niza podataka $x_1, \dots, x_n$.

2. Kolmogorov-Smirnovljev test

Vidjeli smo da empirijska funkcija razdiobe F_n, posebno za velike n, omogućuje dobar uvid u nepoznatu razdiobu vjerojatnosti promatrane slučajne varijable X, pa se odmah nameće ideja da se ona iskoristi pri testiranju hipoteze H_0 da podaci $x_1, \dots, x_n$ potječu od konkretne vjerojatnosne razdiobe kojoj pripada f.r.v. F_0. Promatrajući, naime, empirijsku funkciju razdiobe, može se naslutiti koja je konkretna teorijska vjerojatnosna razdioba, karakterizirana funkcijom F_0, uzrokovala pojavu baš danih mjerenja $x_1, \dots, x_n$. Tu obično pomaže tzv. ***papir vjerojatnosti*** o kojem će biti riječi u idućem poglavlju.

Problem testiranja hipoteza o razdiobi vjerojatnosti razmotren je, doduše, već u IX.3, gdje je opisana primjena hikvadrat-testa, koji je izvorno kreiran za diskretne vjerojatnosne razdiobe, ali se može prilagoditi i za kontinuirane razdiobe. Ovdje će se opisati *Kolmogorov-Smirnovljev test (KS-test)*, koji se primjenjuje samo na kontinuirane razdiobe vjerojatnosti.

KS-test omogućuje testiranje hipoteze H_0: (podaci potječu od kontinuirane vjerojatnosne razdiobe sa f.r.v. F_0), uz zadanu razinu značajnosti α. Kritično područje KS-testa određuje se na temelju test-statistike[2]

[1] U strogom obliku relacija (6) zapisuje se
$$P(\lim_{n\to\infty} \sup_{x\in\mathbf{R}} |\widehat{F}_n(x) - F(x)| = 0) = 1.$$

[2] U formuli (7) korektnije bi bilo pisati sup, umjesto max.

(7) $$D_n = \max_{x\in\mathbf{R}} |F_n(x) - F_0(x)|,$$
kojoj, u uvjetima istinitosti hipoteze H_0, pripada odgovarajuća razdioba vjerojatnosti koja nije ovisna o F_0 nego samo o veličini uzorka n.

Ta se tvrdnja temelji na činjenici da za svaku kontinuiranu slučajnu varijablu X, kojoj pripada f.r.v. F, vrijedi da slučajnoj varijabli $Y = F(X)$ pripada uniformna razdioba $U(0,1)$. Uzme li se, naime, $y \in \langle 0,1\rangle$ i sa G označi f.r.v. slučajne varijable Y, može se pisati

$$G(y) = P(Y \le y) = P(F(X) \le y) = P(X \le F^{-1}(y)) = F[F^{-1}(y)] = y,$$

a to znači da vrijedi

$$G(y) = \begin{cases} 0, & \text{za} \quad y \le 0 \\ y, & \text{za} \quad 0 < y < 1 \\ 1, & \text{za} \quad y \ge 1, \end{cases}$$

odnosno

(8) $$Y = F(X) \sim U(0,1).$$

Transformira li se niz podataka $x_1,\ldots,x_n$ u niz $y_1 = F_0(x_1),\ldots,y_n = F_0(x_n)$ i ako se sa G_n označi empirijska funkcija razdiobe za taj niz, u uvjetima istinitosti hipoteze H_0 odgovarajuća teorijska razdioba je $U(0,1)$. Vrijednost $G_n(y)$ označuje relativnu frekvenciju događaja $\{Y \le y\}$ u nizu $y_1,\ldots,y_n$, a budući da je $y = F_0(x)$ i F_0 strogo monotona funkcija, to je $G_n(y) = F_n(x)$ i $|F_n(x) - F_0(x)| = |G_n(y) - y|$, iz čega proizlazi da je

(9) $$\max_{x\in\mathbf{R}} |F_n(x) - F_0(x)| = \max_{y\in\langle 0,1\rangle} |G_n(y) - y|.$$

Iz (9) se vidi da je maksimalna udaljenost između empirijske funkcije razdiobe za niz podataka $x_1,\ldots,x_n$ i pretpostavljene teorijske funkcije razdiobe vjerojatnosti F_0 jednaka maksimalnoj udaljenosti između empirijske funkcije razdiobe za niz transformiranih podataka $y_1,\ldots,y_n$ i f.r.v. za uniformnu razdiobu vjerojatnosti $U(0,1)$.

Prema tome, da bi se našla razdioba vjerojatnosti statistike D_n iz (7), dovoljno je promotriti $D_n = \max|\widehat{G}_n(y) - y|$, tj. slučaj kada se kao teorijska razdioba uzima $U(0,1)$. Problem određivanja pripadne funkcije razdiobe vjerojatnosti

$$K_n(x) = P_0(D_n \le x), x \in \mathbf{R},$$

vrlo je složen pa se nećemo u to upuštati.

Odmah se vidi da je vrijednost d_n, test-statistike D_n, određeni pokazatelj globalne razlike između empirijske funkcije razdiobe F_n i pretpostavljene teorijske f.r.v. F_0, pa ako se dobije prevelika vrijednost za d_n, onda to indicira da hipotezu H_0 treba odbaciti. To znači da će kritično područje razine značajnosti α biti određeno uvjetom $d_n \ge c_0$, gdje je c_0 određeno tako da vrijedi $P_0(D_n \ge c_0) = \alpha$, odnosno

(10) $$c_0 = K_n^{-1}(1-\alpha),$$

gdje je K_n^{-1} inverzna funkcija od K_n.

Budući da ne postoji jednostavan analitički izraz za funkcije K_n i K_n^{-1}, izrađene su tablice (v. tabl. VIII. u Dodatku) za konkretnu primjenu KS-testa. Kolmogorov je, inače, pokazao da za velike n ($n \to \infty$) slučajnoj varijabli $\sqrt{n}D_n$ pripada funkcija razdiobe vjerojatnosti

(11) $$K(x) = \sum_{k=-\infty}^{\infty} (-1)^k \exp(-2k^2x^2), x > 0,$$

kojom je definirana tzv. *Kolmogorovljeva razdioba*. To praktički znači da se za $n \ge 100$ kritična vrijednost c_0 iz (10) može računati primjenom ove jednostavne tablice

Tablica 1.

α	0,10	0,05	0,01
c_0	$\frac{1{,}22}{\sqrt{n}}$	$\frac{1{,}36}{\sqrt{n}}$	$\frac{1{,}63}{\sqrt{n}}$

Tako, na primjer, dobije li se na uzorku veličine $n = 100$ vrijednost test-statistike $d_n = 0{,}15$, zaključit će se, uz 5% rizika da se odbaci istinita hipoteza, da izmjereni podaci ne potkrepljuju hipotezu o vjerojatnosnoj razdiobi sa f.r.v. F_0. Tada je, naime, $c_0 = \frac{1{,}36}{\sqrt{100}} = 0{,}136$, pa se vidi da točka 0,15 (vrijednost test-statistike) pada u kritično područje $[0{,}136\,;\infty\rangle$.

Glavna prednost KS-testa pred hikvadrat-testom pri testiranju hipoteze o pretpostavljenoj kontinuiranoj razdiobi vjerojatnosti jest ta što on ne zahtijeva grupiranje podataka u razrede, što je inače vrlo proizvoljan postupak kojim se gubi određeni dio informacije o promatranoj pojavi sadržane u danom nizu podataka, a što je glavni nedostatak hikvadrat-testa.

Nedostatak KS-testa očituje se u situaciji kada se najprije, na temelju danih podataka, izvodi procjena parametara pretpostavljene teorijske razdiobe vjerojatnosti, a zatim se na istim podacima primjenjuje i KS-test. Nije, naime, teorijski razjašnjen utjecaj procjene parametara na razdiobu vjerojatnosti test-statistike D_n, što je inače riješeno Fisherovim teoremom za hikvadrat-test.

Što se tiče razlučivosti (v. IX.4) KS-testa, poznato je (v. [4]) da se KS-testom, razine značajnosti α, ne mogu razlučiti vjerojatnosne razdiobe s pripadnim f.r.v. F i G ako je

(12) $$\max_{x\in\mathbf{R}} |F(x) - G(x)| \le \begin{cases} \frac{0{,}84}{\sqrt{n}}, & \text{za} \quad \alpha = 0{,}05 \\ \frac{0{,}65}{\sqrt{n}}, & \text{za} \quad \alpha = 0{,}10\,. \end{cases}$$

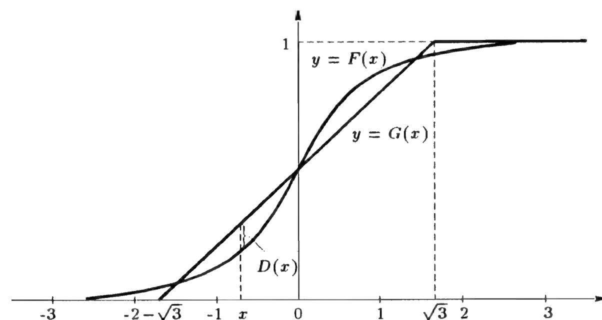
Promotre li se, na primjer, standardna normalna razdioba $N(0,1)$ i uniformna razdioba $U(-\sqrt{3},\sqrt{3})$, tada je

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) d\,t, \quad x \in \mathbf{R},$$

$$G(x) = \begin{cases} 0, & \text{za } x \leq -\sqrt{3} \\ \frac{1}{2}\left(\frac{x}{\sqrt{3}} + 1\right), & \text{za } -\sqrt{3} < x < \sqrt{3} \\ 1, & \text{za } x > \sqrt{3}, \end{cases}$$

tako da je

$$(13)\quad D(x) = F(x) - G(x) = \begin{cases} \Phi(x), & \text{za } x \leq -\sqrt{3} \\ \Phi(x) - \frac{1}{2}\left(\frac{x}{\sqrt{3}} + 1\right), & \text{za } -\sqrt{3} < x < \sqrt{3} \\ 1 - \Phi(x), & \text{za } x \geq \sqrt{3}. \end{cases}$$



Slika 28. Odnos f.r.v. za N(0,1) i U($-\sqrt{3}, \sqrt{3}$)

Deriviranjem (13) po x dobiva se

$$D'(x) = \begin{cases} \varphi(x), & \text{za } x \leq -\sqrt{3} \\ \varphi(x) - \frac{1}{2\sqrt{3}}, & \text{za } -\sqrt{3} < x < \sqrt{3} \\ -\varphi(x), & \text{za } x \geq \sqrt{3}, \end{cases}$$

gdje je

$$\Phi'(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), x \in \mathbf{R},$$

f.g.v. za N(0, 1).

Rješavanjem jednadžbe $D'(x) = 0$ po x dobivaju se rješenja $x_1 = 0{,}8$ i $x_2 = -0{,}8$, iz čega se zaključuje da je

$$\max_{x \in \mathbf{R}} |D(x)| = |F(0{,}8) - G(0{,}8)| = |F(-0{,}8) - G(-0{,}8)| \approx 0{,}057.$$

Može se reći da najveća udaljenost između F i G približno iznosi 0,057, pa se u vezi sa (12) odmah možemo pitati za koje veličine uzorka n nije moguće razlučiti

spomenute razdiobe vjerojatnosti, recimo, uz razinu značajnosti $\alpha = 0{,}05$? Odgovor se dobiva primjenom relacije (12), iz koje proizlazi

$$0{,}057 \leq \frac{0{,}84}{\sqrt{n}} \Rightarrow n \leq 218.$$

Prema tome, s manje od 218 podataka nije moguće KS-testom, uz razinu značajnosti od 5%, razlučiti standardnu normalnu razdiobu N(0, 1) od uniformne razdiobe U($-\sqrt{3}, \sqrt{3}$).

Općenito se može tvrditi da za pouzdani zaključak pri primjeni KS-testa treba raspolagati s vrlo velikim uzorkom. Ako se s relativno malim uzorkom donese odluka o prihvaćanju hipoteze H_0, ranija nam razmatranja pokazuju da pretpostavljenu razdiobu (F_0) treba s rezervom smatrati stvarnom teorijskom razdiobom kojoj se podvrgavaju dani podaci, jer bi KS-testom bile prihvatljive i mnoge druge vjerojatnosne razdiobe.

Praktična primjena KS-testa, a slično je i s hikvadrat-testom, kada se ne raspolaže s dovoljno velikim uzorcima, opravdava se činjenicom da stvarne vjerojatnosne razdiobe koje opisuju statističke zakonitosti realnih slučajnih fenomena nisu, zapravo, ni normalne, ni uniformne, ni eksponencijalne itd., već su samo približno normalne, približno uniformne, približno eksponencijalne itd., pa je riječ o tome da se, na bar donekle objektivan način, utvrdi koji od matematičkih modela dolazi, odnosno ne dolazi, u obzir za opisivanje promatranoga realnog statističkog fenomena.

3. Papir vjerojatnosti

Jedno od važnih praktičnih pitanja u problemu prilagodbe teorijske razdiobe vjerojatnosti danim statističkim podacima svakako je pitanje kako naslutiti, odnosno pogoditi, teorijsku razdiobu koja će se dobro suglasiti s danim nizom podataka. Ako su posrijedi podaci dobiveni mjerenjem diskretne slučajne varijable X i broj n podataka dovoljno velik, grafikon relativnih frekvencija vrlo zorno upućuje na određeni tip vjerojatnosne razdiobe.

Kada je riječ o podacima dobivenim mjerenjem kontinuirane slučajne varijable X, onda se stvari kompliciraju. Za izradbu grafikona relativnih frekvencija nužno je grupiranje podataka u razrede, a vidjeli smo (v. I.6, sl. 8. i 10) da izgled toga grafikona bitno ovisi o načinu grupiranja (širina razreda, broj razreda i dr.). Stoga se ne možemo pouzdano osloniti na grafikon relativnih frekvencija kao siguran putokaz za pretpostavku o teorijskoj razdiobi.

Kao druga mogućnost ostaje empirijska funkcija razdiobe i njen graf, gdje ne dolazi do gubitka informacije zbog grupiranja u razrede, ali je očigledno da razlike u grafovima funkcija razdiobe vjerojatnosti za različite tipove teorijskih razdioba nisu toliko tipične da bi se lako uočile u pravokutnom koordinatnom sustavu s linearnim ljestvicama na koordinatnim osima. Zato se nameće pomisao da se organizira koordinatni sustav s takvim ljestvicama na koordinatnim osima, koji će omogućiti jasno prepoznavanje vjerojatnosne razdiobe iz koje potječu, odnosno ne potječu, dani statistički podaci.

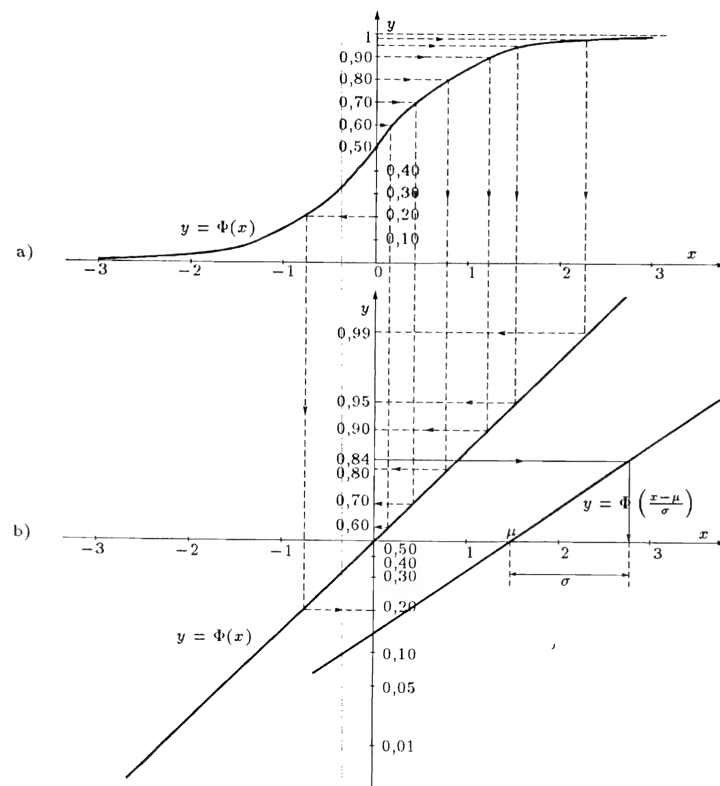
Bit ideje je u tome da se izaberu takve ljestvice na koordinatnim osima, uz koje

će graf f.r.v. za određenu klasu vjerojatnosnih razdioba, recimo za sve normalne razdiobe, biti pravac. Ucrta li se u taj koordinatni sustav empirijska funkcija razdiobe i ako podaci stvarno potječu od neke teorijske razdiobe iz dotične klase, njezin graf se neće mnogo razlikovati od pravca. Ako se on znatno razlikuje od pravca, pokušat će se prikazati u koordinatnom sustavu s ljestvicama na koordinatnim osima prilagođenim nekoj drugoj klasi vjerojatnosnih razdioba. Takav koordinatni sustav općenito se zove *papir vjerojatnosti*.

Papiri vjerojatnosti za pojedine klase vjerojatnosnih razdioba, slično kao i tablice, unaprijed su načinjeni i mogu se obično nabaviti kao i svaki drugi tiskani materijal. Tako postoji tzv. *normalni papir vjerojatnosti*, koji je pripremljen za klasu normalnih razdioba $N(\mu, \sigma^2)$, i slični za druge klase razdioba.

Postupak izrade normalnog papira vjerojatnosti ilustriran je na sl. 29. Na sl. 29a nacrtan je graf f.r.v. za standardnu normalnu razdiobu $N(0,1)$ u pravokutnom koordinatnom sustavu s linearnim ljestvicama na apscisnoj i ordinatnoj osi.

Da bi se dobila sl. 29b najprije se nacrtaju koordinatne osi i za ishodište uzme točka $(0\,;\,0{,}50)$. Na apscisnoj osi zadržava se ista linearna ljestvica kao i na sl. 29a, dok se ljestvica na ordinatnoj osi dobiva tako da se najprije nacrta proizvoljan kosi



Slika 29. Skica postupka izrade normalnog papira vjerojatnosti

pravac (recimo pod kutom od 45° prema apscisnoj osi) kroz ishodište, a zatim se linearna ljestvica na ordinatnoj osi sl. 29a, pomoću krivulje $y = \Phi(x)$ i nacrtanog pravca, preslika na ordinatnu os sl. 29b, kao što je opisano pomoću strelica na sl. 29. Time je dobivena nelinearna ljestvica na ordinatnoj osi sl. 29b i u tom je koordinatnom sustavu graf f.r.v. standardne normalne razdiobe $N(0,1)$ baš nacrtani pravac.

Zamislimo sada koordinatni sustav na sl. 29b bez ucrtanog pravca i to je onda normalni papir vjerojatnosti. Graf f.r.v. za normalnu razdiobu $N(\mu, \sigma^2)$ $(\mu \in \mathbf{R}, \sigma > 0)$ u tom će koordinatnom sustavu biti pravac koji prolazi točkom $(\mu; 0{,}50)$ i točkom $(\mu+\sigma; 0{,}84)$, što proizlazi iz osnovnih svojstava normalne razdiobe

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right), \quad F(\mu) = \Phi(0) = 0{,}50\,, \quad F(\mu+\sigma) = \Phi(1) \approx 0{,}84\,.$$

Praktičnu primjenu normalnog papira vjerojatnosti ilustrirat ćemo idućim primjerom.

1. primjer

Uzmimo prvih deset podataka iz 5. primjera u I.4 (tlačna čvrstoća betonskih kocki) i odmah ih poredajmo po veličini:

$$x_1 = 30{,}97,\quad x_2 = 33{,}56,\quad x_3 = 34{,}47,\quad x_4 = 35{,}76,\quad x_5 = 38{,}79,$$
$$x_6 = 39{,}22,\quad x_7 = 40{,}15,\quad x_8 = 42{,}63,\quad x_9 = 45{,}00,\quad x_{10} = 47{,}12.$$

Očigledno je da se u dani normalni papir vjerojatnosti ne može ucrtati empirijska funkcija razdiobe za navedene podatke, pa ih treba transformirati pomoću afine transformacije, tako da to bude moguće učiniti. U tu svrhu uočimo najmanju ($x_{\min} = 30{,}97$) i najveću ($x_{\max} = 47{,}12$) vrijednost u danom nizu podataka i načinimo

$$\frac{x_{\max} - x_{\min}}{6} = 2{,}69\,, \qquad \frac{x_{\min} + x_{\max}}{2} = 39{,}04.$$

Radi lakšeg računanja uzmimo $a = 3$ (umjesto 2,69) i $b = 40$ (umjesto 39,04) i provedimo transformaciju danih podataka (x) u novi niz podataka (z) pomoću formule

$$z = \frac{x-b}{a} = \frac{x-40}{3}.$$

Dobiva se niz

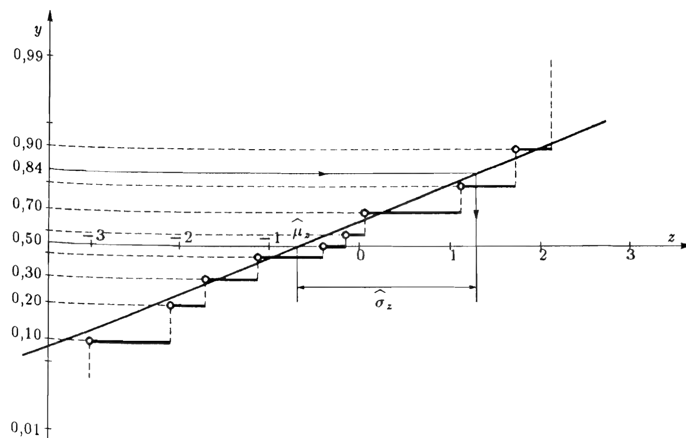
$$z_1 = -3{,}01,\quad z_2 = -2{,}15,\quad z_3 = -1{,}84,\quad z_4 = -1{,}41,\quad z_5 = -0{,}40,$$
$$z_6 = -0{,}26,\quad z_7 = 0{,}05,\quad z_8 = 0{,}88,\quad z_9 = 1{,}67,\quad z_{10} = 2{,}37.$$

Sada se u normalni papir vjerojatnosti mogu ucrtati točke

$$(-3{,}01\,;\,0{,}10),\quad (-2{,}15\,;\,0{,}20),\quad (-1{,}84\,;\,0{,}30),\quad (-1{,}41\,;\,0{,}40),$$
$$(-0{,}40\,;\,0{,}50),\quad (-0{,}26\,;\,0{,}60),\quad (0{,}05\,;\,0{,}70),\quad (0{,}88\,;\,0{,}80),$$
$$(1{,}67\,;\,0{,}90).$$

Točka (2,37 ; 1,00) očigledno se ne može ucrtati u normalni papir vjerojatnosti, jer se ona zapravo nalazi u beskonačnosti.

Navedenim točkama određen je graf empirijske funkcije razdiobe za niz podataka o slučajnoj varijabli $Z = \frac{X-b}{a}$, pa se vidi da se dobiveni niz točaka može vrlo dobro aproksimirati pravcem, što upućuje na zaključak da taj niz podataka $(z_1, \ldots, z_{10})$ potječe od normalne razdiobe $N(\mu_z, \sigma_z^2)$. Budući da je $X = aZ + b$, slijedi da niz izvornih podataka $x_1, \ldots, x_{10}$ potječe od normalne razdiobe $N(a\mu_z + b, a^2\sigma_z^2)$.



Slika 30. Empirijska funkcija razdiobe za podatke iz 1. primjera

Sa sl. 30. mogu se "očitati" i procjene $\widehat{\mu}_z$ i $\widehat{\sigma}_z$ za nepoznate parametre μ_z i σ_z, pa se $\widehat{\mu} = a\widehat{\mu}_z + b$ i $\widehat{\sigma} = a\widehat{\sigma}_z$ mogu uzeti kao procjene za nepoznate parametre μ i σ normalne razdiobe $N(\mu, \sigma^2)$, od koje se pretpostavlja da potječu izvorni podaci $x_1, \ldots, x_{10}$.

U promatranom primjeru (v. sl. 30) nalazimo da je $\widehat{\mu}_z = -0{,}75$ i $\widehat{\sigma}_z = 1{,}90$, pa je $\widehat{\mu} = 3\,(-0{,}75) + 40 = 37{,}75$ i $\widehat{\sigma} = 3 \cdot 1{,}90 = 5{,}70$.

Jasno je da tako dobivene procjene za nepoznate parametre nisu egzaktne, jer se oslanjaju na subjektivni postupak "očitavanja" brojčanih vrijednosti s dobivenoga grafikona. No, to i nije bila glavna svrha primjene papira vjerojatnosti, jer za procjene nepoznatih parametara postoje egzaktne metode opisane u VI. poglavlju.

Kao što je već ranije rečeno, papir vjerojatnosti omogućuje jednostavan, jasan i pregledan uvid u eventualnu usklađenost, odnosno neusklađenost, empirijske funkcije razdiobe s pretpostavljenom klasom teorijskih razdioba vjerojatnosti i to je njegova osnovna svrha.

Primjedba

Na kraju I. i II. poglavlja primijećeno je da primjena računskih strojeva (osobnih računala) omogućuje brzo i točno provođenje statističkih proračuna i grafičko prikazivanje statističkih podataka. Statističkim programskim paketima redovito je obuhvaćeno oblikovanje koordinatnih sustava s različitim funkcijskim ljestvicama i prikazivanje statističkih podataka u njima. Time je automatizirana izradba "papira vjerojatnosti" i njihova primjena.

Zadaci

1. Skicirajte graf odgovarajuće empirijske funkcije razdiobe za podatke o diskretnome statističkom obilježju X iz zad. 1-7. u I. poglavlju.
2. Skicirajte graf empirijske funkcije razdiobe za prvih dvadeset podataka u 5. primjeru u I.4.
3. Skicirajte graf empirijske funkcije razdiobe za podatke o kontinuiranome statističkom obilježju X iz zad. 8-14. u I. poglavlju.
4. Primjenom KS-testa, uz razinu značajnosti $\alpha = 0{,}05$, testirajte hipotezu da podaci iz
   a) zad. 8. u I. pogl. potječu od $N(300, 289)$,
   b) zad. 9. u I. pogl. potječu od $Ex(0{,}005)$,
   c) zad. 11. u I. pogl. potječu od $N(3{,}5\,;\,0{,}5)$,
   d) zad. 14. u I. pogl. potječu od $U(0\,;\,6{,}5)$.
5. Nađite najveću udaljenost između f.r.v. standardne normalne razdiobe $N(0, 1)$ i Laplaceove razdiobe parametra $\alpha = \sqrt{2}$. Primjenom formule (12) odredite veličinu uzorka n_0 ispod koje nije moguće razlučiti te dvije razdiobe primjenom KS-testa, uz razinu značajnosti:
   a) $\alpha = 0{,}05$ b) $\alpha = 0{,}10$.
6. Koliko je najmanje (n_0) podataka potrebno da bi se razlučila standardna normalna razdioba $N(0, 1)$ od tzv. *pomućene normalne razdiobe (contaminated normal distribution)* kojoj pripada f.r.v.

$$G(x) = p\Phi(x) + (1-p)\Phi\left(\frac{x}{\sigma}\right), x \in \mathbf{R},$$

   gdje je $0 < p < 1$ i $\sigma > 1$? Nađite konkretno n_0 za $p = 0{,}9$ i $\sigma = 2$.
7. Što je papir vjerojatnosti za klasu uniformnih razdioba? Može li se, uvidom u empirijsku funkciju razdiobe za podatke iz zad. 14. u I. poglavlju, zaključiti da se dani podaci pokoravaju nekoj uniformnoj razdiobi?
8. Transformirajte podatke iz zad. 8. u I. poglavlju na odgovarajući način i zatim u normalni papir vjerojatnosti ucrtajte empirijsku funkciju razdiobe za transformirane podatke. Povucite pravac koji najbolje aproksimira tu funkciju, "očitajte" procjene za očekivanje (μ_z) i standardnu devijaciju (σ_z), a zatim

izračunajte odgovarajuće procjene za parametre μ i σ normalne razdiobe za koju se pretpostavlja da pripada izvornim podacima. Usporedite dobivene vrijednosti s vrijednostima procjena dobivenim na drugi način.

9. Konstruirajte papir vjerojatnosti za klasu eksponencijalnih razdioba $\text{Ex}(\alpha)$ ($\alpha > 0$). (Uputa: Uzmite $\text{Ex}(1)$ kao standardnu eksponencijalnu razdiobu. Njezin graf nad intervalom $[0, \infty)$ preslikat će se u polupravac iz ishodišta u koordinatnom sustavu s linearnom ljestvicom na apscisnoj osi i funkcijskom ljestvicom na ordinatnoj osi dobivenom na način opisan na sl. 29.)

10. Ucrtajte u papir vjerojatnosti za eksponencijalne razdiobe graf empirijske funkcije razdiobe niza podataka dobivenih prikladnom transformacijom podataka iz zad. 9. u I. poglavlju. (Uputa: Provedite transformaciju po formuli $z = \dfrac{x}{a}$, gdje je za a uzet broj približno jednak aritmetičkoj sredini danih podataka.) Može li se zaključiti da se dani podaci podvrgavaju nekoj eksponencijalnoj razdiobi? Kako biste s dobivenog grafikona "odčitali" procjenu za nepoznati parametar eksponencijalne razdiobe?

XI. Regresijska analiza

1. Regresijska zavisnost

U prethodnim poglavljima razmotreni su različiti matematički modeli teorije statističkog zaključivanja u kojima je slučajni uzorak interpretiran kao niz ponovljenih nezavisnih mjerenja jedne te iste slučajne varijable, odnosno u nekim situacijama i dvodimenzinalnoga slučajnog vektora. Mnogi praktični problemi, međutim, zahtijevaju da se donesu određeni zaključci o nizu slučajnih varijabli $Y_1, \ldots, Y_n$, koje ovise o neslučajnoj (nezavisnoj) varijabli x, na temelju niza sparenih mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$, gdje su $x_1, \ldots, x_n$ vrijednosti nezavisne varijable x, a $y_1, \ldots, y_n$ su odgovarajuće vrijednosti slučajnih varijabli $Y_1, \ldots, Y_n$.

Tako, na primjer, x može označivati dob, a Y krvni tlak osobe, pa se postavlja zadatak da se istraži veza između krvnog tlaka i dobi. Poznato je, naime, da je krvni tlak za osobe iste dobi slučajna veličina, za koju se može pretpostaviti određena statistička zakonitost, tj. može se modelirati kao slučajna varijabla s pripadnom razdiobom vjerojatnosti. Uzme se, recimo, normalna razdioba $\text{N}(\mu, \sigma^2)$.

Također je poznato da se sa starenjem povećava krvni tlak, pa se prirodno nameće zadatak da se istraži i matematički opiše statistička zakonitost koja obuhvaća i vremensku promjenljivost krvnog tlaka. To znači da se mora promatrati jedna familija normalnih razdioba, tako da svakoj dobi x pripada odgovarajuća normalna razdioba $\text{N}(\mu(x), \sigma^2(x))$ krvnog tlaka Y_x. Činjenica da se sa starenjem povećava krvni tlak odrazit će se na funkciju $x \mapsto \mu(x)$, koja označuje srednju vrijednost krvnog tlaka osobe dobi x, tako da će ta funkcija monotono rasti, dok će, recimo, $\sigma^2(x) = \sigma^2$ biti neovisno o x.

Općenito se zadatak svodi na to da se ustanovi priroda ovisnosti slučajnih varijabli Y_x o nezavisnoj varijabli x, na temelju niza sparenih mjerenja (x_i, y_i), $(i = 1, \ldots, n)$.

Ako se matematički model definira relacijom

$$Y_i = \mu_t(x_i) + \mathcal{E}_i, \qquad i = 1, \ldots, n, \tag{1}$$

gdje je $x \mapsto \mu_t(x)$ realna funkcija jedne realne varijable određena parametrom t ($t \in \Theta$), a $\mathcal{E}_1, \ldots, \mathcal{E}_n$ nezavisne slučajne varijable s očekivanjem $\text{E}[\mathcal{E}_i] = 0$ i varijancom $\text{V}[\mathcal{E}_i] = \sigma^2$, onda se govori o ***jednodimenzionalnome regresijskom modelu***.

Može se reći da se u opisanome regresijskom modelu pretpostavlja da je rezultat mjerenja promatrane pojave u momentu x_i slučajna varijabla Y_i, koja je nastala zbrajanjem vrijednosti ***regresijske funkcije*** $\mu_t(x_i)$ i slučajne greške $\mathcal{E}_i$.

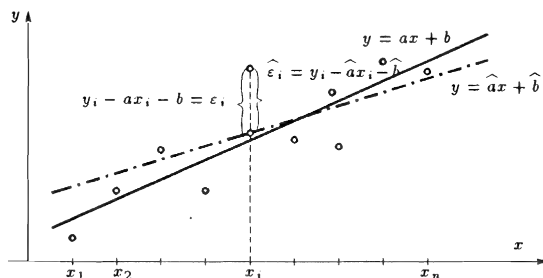
Osnovni je zadatak regresijske analize da se, na temelju niza sparenih mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$, procijene nepoznati parametri t i σ^2 (t može biti, i re-

dovito i jest, vektorski parametar). Preciznije govoreći, riječ je o tome da se definiraju odgovarajući procjenitelji za nepoznate parametre t i σ^2, polazeći od $((x_1, Y_1), \dots, (x_n, Y_n))$ kao slučajnog uzorka. Primijetimo da se ovdje slučajni uzorak shvaća u nešto drukčijem smislu nego ranije, jer su $Y_1, \dots, Y_n$, doduše, međusobno nezavisne slučajne varijable, ali ne nužno i sa zajedničkom razdiobom vjerojatnosti.

Da bi se mogla postaviti što realnija pretpostavka o regresijskoj funkciji, obično se podaci $(x_1, y_1), \dots, (x_n, y_n)$ prikazuju kao točke u pravokutnome koordinatnom sustavu u ravnini, iz čega se dade naslutiti tip zakonitosti koji povezuje nezavisnu varijablu x i vrijednosti slučajnih varijabli Y_x. Upućuju li dobivene točke na aproksimaciju pravcem, uzet će se $\boldsymbol{t} = (a, b) \in \mathbf{R}^2$ i staviti

(2) $$\mu_{\boldsymbol{t}}(x) = ax + b,$$

tj. pretpostavit će se da je regresijska funkcija polinom prvog stupnja, odnosno da je regresijska linija pravac. Parametar a zove se u tom slučaju *koeficijent regresije*, a pravac $y = ax + b$ zove se *regresijski pravac*.



Slika 31. Skica regresijskog problema

Problem regresijske analize sastoji se u određivanju "dobrih" procjena $\widehat{a}$, $\widehat{b}$ i $\widehat{\sigma}^2$ za nepoznate parametre a, b i σ^2. Odmah se postavlja pitanje može li neka od već razmotrenih metoda (ML, metoda momenata i dr.) za dobivanje procjenitelja poslužiti da se i u ovom slučaju odrede prikladni procjenitelji nepoznatih parametara. Ako se ne uvedu dodatne pretpostavke o tipu vjerojatnosne razdiobe slučajnih varijabli $\mathcal{E}_i$, onda je očigledno da nijedna od metoda opisanih u VII. poglavlju ne dolazi u obzir za procjenu parametara u danom problemu regresijske analize.

Za procjenu parametara u problemima regresijske analize redovito se primjenjuje ***metoda najmanjih kvadrata*** (v. III.4). U promatranom slučaju to znači da se procjene $\widehat{a}$ i $\widehat{b}$ trebaju odrediti tako da vrijedi

(3) $$\min_{(a,b)\in\mathbf{R}^2} \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (y_i - \widehat{a}x_i - \widehat{b})^2.$$

Ako je $n > 2$ i svi x_i $(i = 1, \dots, n)$ nisu međusobno jednaki, onda se lako pokazuje da sustav jednadžbi

$$\frac{\partial}{\partial a}\left[\sum_{i=1}^{n}(y_i - ax_i - b)^2\right] = -2\sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0$$

$$\frac{\partial}{\partial b}\left[\sum_{i=1}^{n}(y_i - ax_i - b)^2\right] = -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0$$

ima jednoznačno rješenje

(4) $$a = \widehat{a} = \frac{s_{xy}}{s_x^2} = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})y_i,$$

(5) $$b = \widehat{b} = \overline{y} - \widehat{a}\,\overline{x} = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{\overline{x}}{s_x^2}(x_i - \overline{x})\right] y_i,$$

gdje je

(6) $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$
$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Stavi li se

(7) $$\widehat{A} = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \overline{x})Y_i,$$

(8) $$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \frac{\overline{x}}{s_x^2}(x_i - \overline{x})\right] Y_i,$$

(9) $$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{A}x_i - \widehat{B})^2,$$

slučajne varijable $\widehat{A}$, $\widehat{B}$ i S^2, kao određene funkcije slučajnog uzorka $(x_1, Y_1), \dots, (x_n, Y_n)$, mogu se smatrati procjeniteljima nepoznatih parametara a, b i σ^2. Kaže se da su to ***procjenitelji u smislu metode najmanjih kvadrata*** ili kraće ***MNK-procjenitelji***. Za pravac $y = \widehat{a}x + \widehat{b}$ kaže se da je procjena za nepoznati regresijski pravac $y = ax + b$, dobivena metodom najmanjih kvadrata na temelju konkretnog niza mjerenja $(x_1, y_1), \dots, (x_n, y_n)$.

Često se, međutim, i pravac $y = \widehat{a}x + \widehat{b}$ zove regresijski pravac.

Primijetimo da iz (5) proizlazi

$$\overline{y} = \widehat{a}\,\overline{x} + \widehat{b},$$

što pokazuje da regresijski pravac prolazi točkom $(\overline{x}, \overline{y})$, koja se može interpretirati kao određeno središte izmjerenih podataka.

Uzmu li se u obzir pretpostavke modela (1) i (2), odmah se vidi da je

$$(10)\qquad \mathrm{E}[Y_i] = \mu_t(x_i) = ax_i + b, \quad \mathrm{V}[Y_i] = \mathrm{V}[\mathcal{E}_i] = \sigma^2, \qquad i = 1, \ldots, n,$$

iz čega proizlazi da je

$$(11)\qquad \mathrm{E}[\widehat{A}] = a, \quad \mathrm{V}[\widehat{A}] = \frac{\sigma^2}{ns_x^2},$$

$$(12)\qquad \mathrm{E}[\widehat{B}] = b, \quad \mathrm{V}[\widehat{B}] = \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right),$$

$$(13)\qquad \mathrm{E}[S^2] = \sigma^2.$$

To pokazuje da su $\widehat{A}$, $\widehat{B}$ i S^2 nepristrani procjenitelji za parametre a, b i σ^2. Za $\widehat{A}$ i $\widehat{B}$ se vidi da su i konzistentni procjenitelji, a može se dokazati da su također i asimptotski normalni procjenitelji, što znači da se za velike n može smatrati da

$$(14)\qquad \widehat{A} \sim \mathrm{N}\left(a, \frac{\sigma^2}{ns_x^2}\right),$$

$$(15)\qquad \widehat{B} \sim \mathrm{N}\left(b, \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right)\right).$$

To omogućuje da se, primjenom metode opisane u VII.3, odrede intervali povjerenja zadane pouzdanosti γ, za nepoznate parametre a i b, pri čemu se, umjesto nepoznatog parametra σ^2, uzima vrijednost s^2 nepristranog procjenitelja S^2 iz (9). Odmah se vidi da je

$$(16)\qquad s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{a}x_i - \widehat{b})^2 = \frac{n}{n-2}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right),$$

gdje je

$$(17)\qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Sada se, za svaki $x \in \mathbf{R}$, može promatrati i statistika $\widehat{A}x + \widehat{B}$, kao nepristran, konzistentan i asimptotski normalan procjenitelj za $ax + b$, te odgovarajući interval povjerenja zadane pouzdanosti γ. Očigledno je, naime, $\mathrm{E}[\widehat{A}x + \widehat{B}] = x\mathrm{E}[\widehat{A}] + \mathrm{E}[\widehat{B}] = ax + b$, što pokazuje da je riječ o nepristranom procjenitelju. Slučajne varijable $\widehat{A}$ i $\widehat{B}$ općenito nisu nekorelirane, pa se, na temelju (7) i (8), lako izračunava njihova kovarijanca

$$(18)\qquad \mathrm{Cov}(\widehat{A}, \widehat{B}) = \mathrm{E}[\widehat{A}\,\widehat{B}] - \mathrm{E}[\widehat{A}]\mathrm{E}[\widehat{B}] = -\frac{\sigma^2\overline{x}}{ns_x^2},$$

a zatim se iz (11), (12) i (18) dobiva

$$(19)\qquad \mathrm{V}[\widehat{A}x + \widehat{B}] = x^2\mathrm{V}[\widehat{A}] + \mathrm{V}[\widehat{B}] + 2x\mathrm{Cov}(\widehat{A}, \widehat{B}) = \frac{\sigma^2}{n}\left[1 + \frac{(x - \overline{x})^2}{s_x^2}\right],$$

iz čega se razabire da je riječ o konzistentnom procjenitelju.

Stavljajući $\widehat{T} = \widehat{A}x + \widehat{B}$, može se primijeniti metoda opisana u VII.3. za određivanje intervala povjerenja pouzdanosti γ pri velikim uzorcima, za nepoznatu vrijednost $ax + b$. Na temelju formula (38) iz VII.3, gdje treba, umjesto $R_n(\widehat{T})$, staviti $\mathrm{V}[\widehat{A}x + \widehat{B}]$ iz (19), dobivaju se vrijednosti donjeg i gornjeg ruba intervala povjerenja

$$(20)\qquad g_1(x) = \widehat{a}x + \widehat{b} - \frac{d(x)}{2}, \quad g_2(x) = \widehat{a}x + \widehat{b} + \frac{d(x)}{2},$$

gdje je

$$(21)\qquad d(x) = 2z_\gamma\frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{(x - \overline{x})^2}{s_x^2}}$$

širina intervala povjerenja na mjestu $x \in \mathbf{R}$.

Iz (21) se razabire da širina intervala povjerenja ovisi, osim o pouzdanosti γ, veličini uzorka n i karakteristici greške σ^2, još i o vrijednosti nezavisne varijable x, tako da je širina intervala najmanja za $x = \overline{x}$. Također se vidi da veće rasipanje nezavisne varijable (s_x^2) utječe na sužavanje intervala povjerenja.

1. primjer

Na skupu od $n = 100$ osoba različite dobi mjeren je krvni tlak (sistolički tlak). Nakon statističke obrade izmjerenih podataka dobivene su ove vrijednosti relevantnih veličina:

$\overline{x} = 45$ (prosječna dob osoba u godinama)
$s_x = 12$ (standardna devijacija dobi)
$\overline{y} = 130$ (prosječni sistolički tlak)
$s_y = 9$ (standardna devijacija sistoličkog tlaka)
$s_{xy} = 86$.

Na temelju formula (4) i (5) dobivaju se procjene

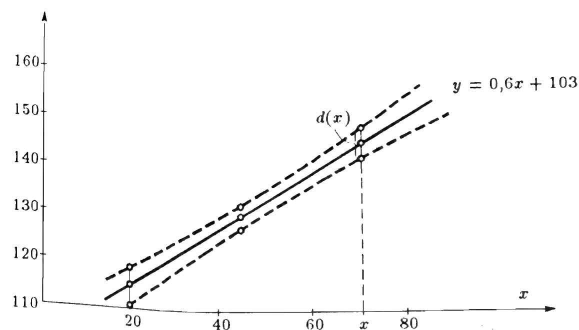
$$\widehat{a} = \frac{s_{xy}}{s_x^2} = \frac{86}{144} \approx 0{,}6\,, \quad \widehat{b} = \overline{y} - \widehat{a}\overline{x} = 130 - 0{,}6 \cdot 45 = 103,$$

nepoznatih parametara a i b regresijskog pravca, što omogućuje da se napiše njegova jednadžba

$$y = 0{,}6x + 103$$

i nacrta njegov graf (sl. 32).

Budući da je veličina uzorka dovoljno velika ($n = 100$), može se, primjenom formule (16), uzeti da je $\sigma^2 \approx s^2 = \dfrac{100}{98}\left(81 - \dfrac{7\,396}{144}\right) \approx 31$, odnosno $\sigma \approx 5{,}7$. Uzme li se $\gamma = 0{,}95$, iz (21) se dobiva širina intervala povjerenja za nepoznatu veličinu $ax + b$, izražena u obliku



Slika 32. Prikaz regresijske ovisnosti i 95-postotnog intervala povjerenja

$$d(x) = 2 \cdot 1{,}96 \cdot \frac{5{,}7}{10}\sqrt{1 + \frac{(x-45)^2}{144}} = 0{,}19\sqrt{144 + (x-45)^2}.$$

Tako se, za $x = 45$, dobiva najmanja vrijednost $d(45) = 2{,}23$, dok se za $x = 20$ dobiva $d(20) = 5{,}26$, isto kao i za $x = 70$.

2. Linearna regresija

Bitna je pretpostavka, koja je omogućila dobivanje formula (14), (15), (20) i (21), da se raspolaže s velikim uzorkom ($n \to \infty$), jer se tada može iskoristiti asimptotska normalnost procjenitelja $\widehat{A}$ i $\widehat{B}$ pri određivanju odgovarajućih intervala povjerenja. U mnogim praktičnim problemima, međutim, neće se raspolagati s dovoljno velikim brojem mjerenja, pa se moraju prihvatiti neke druge pretpostavke, koje će omogućiti određivanje vjerojatnosnih razdioba procjenitelja $\widehat{A}$, $\widehat{B}$ i S^2 za nepoznate parametre a, b i σ^2 regresijskog modela

(22) $$Y_i = ax_i + b + \mathcal{E}_i, \quad i = 1, \ldots, n.$$

To će se postići usvajanjem dodatne pretpostavke da nezavisnim slučajnim varijablama $\mathcal{E}_1, \ldots, \mathcal{E}_n$ pripada zajednička normalna razdioba $\mathrm{N}(0, \sigma^2)$. Pretpostavlja se, znači, da je izmjerena vrijednost y_i ($i = 1, \ldots, n$), koja se zove još i *izlaz (output)*, posljedica djelovanja funkcijske (afine) ovisnosti $x_i \mapsto ax_i + b$ i normalne slučajne greške $\mathcal{E}_i$ s očekivanjem nula i standardnom devijacijom $\sigma > 0$.

Iz (22) se razabire da su $Y_1, \ldots, Y_n$ nezavisne slučajne varijable i da vrijedi

(23) $$Y_i \sim \mathrm{N}(ax_i + b, \sigma^2), \quad i = 1, \ldots, n,$$

a iz (7) i (8) da su $\widehat{A}$ i $\widehat{B}$ linearne kombinacije slučajnih varijabli $Y_1, \ldots, Y_n$, pa na temelju izričaja točke 1. u V.6, proizlazi

(24) $$\widehat{A} \sim \mathrm{N}\left(a, \frac{\sigma^2}{ns_x^2}\right),$$

(25) $$\widehat{B} \sim \mathrm{N}\left(b, \frac{\sigma^2}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right)\right).$$

Relacijama (24) i (25) izriče se formalno isto što i relacijama (14) i (15), uz primjedbu da u upravo opisanom modelu regresije one vrijede za svako $n > 2$, dok u modelu iz XI.1. vrijede samo za velike n (praktički za $n > 100$).

Ostalo je još da se u ovom modelu razjasni situacija oko razdiobe vjerojatnosti slučajne varijable S^2 iz (9). Vidi se da je S^2 izraženo kao zbroj kvadrata određene linearne kombinacije normalnih slučajnih varijabli Y_i, $\widehat{A}$ i $\widehat{B}$, koje međutim nisu nezavisne, što onemogućuje neposrednu primjenu izričaja točke 5. iz V.6. To, ipak, upućuje na zaključak da će slučajnoj varijabli S^2, kao zbroju kvadrata normalnih slučajnih varijabli, pripadati vjerojatnosna razdioba povezana na određeni način s hikvadrat-razdiobom. To se zaista može dokazati (v. XII.6), tako da vrijedi

(26) $$\frac{n-2}{\sigma^2}S^2 \sim \chi^2(n-2),$$

kao i činjenica da su $\widehat{A}$ i S^2, te $\widehat{B}$ i S^2 nezavisne slučajne varijable.

Relacije (24), (25) i (26) omogućuju da se egzaktno odrede intervali povjerenja zadane pouzdanosti γ za nepoznate parametre a i b, te za vrijednost $ax + b$ regresijske funkcije za svaki $x \in \mathbf{R}$. Pozivajući se, naime, na točku 7. iz V.6, zaključuje se da

(27) $$T_1 = \frac{\dfrac{\widehat{A} - a}{\dfrac{\sigma}{s_x\sqrt{n}}}}{\sqrt{\dfrac{\dfrac{n-2}{\sigma^2}S^2}{n-2}}} = \frac{\widehat{A} - a}{S}s_x\sqrt{n} \sim \mathrm{t}(n-2),$$

(28) $$T_2 = \frac{\dfrac{\widehat{B} - b}{\dfrac{\sigma}{\sqrt{n}}\sqrt{1 + \dfrac{\overline{x}^2}{s_x^2}}}}{\sqrt{\dfrac{\dfrac{n-2}{\sigma^2}S^2}{n-2}}} = \frac{\widehat{B} - b}{S}\sqrt{\frac{n}{1 + \dfrac{\overline{x}^2}{s_x^2}}} \sim \mathrm{t}(n-2).$$

To omogućuje da se odrede rubovi odgovarajućih intervala povjerenja zadane pouzdanosti γ (v. VII.2). Za parametar a dobiva se

(29) $$g_1 = \widehat{a} - \tau_\gamma \frac{s}{s_x\sqrt{n}}, \qquad g_2 = \widehat{a} + \tau_\gamma \frac{s}{s_x\sqrt{n}},$$

a za parametar b dobiva se

(30) $$g_1 = \widehat{b} - \tau_\gamma \frac{s}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{s_x^2}}, \qquad g_2 = \widehat{b} + \tau_\gamma \frac{s}{\sqrt{n}}\sqrt{1 + \frac{\overline{x}^2}{s_x^2}}.$$

Da bismo odredili interval povjerenja za vrijednost regresijske funkcije $\mu_t = ax + b$, primijetimo najprije da iz (8) proizlazi

(31) $$\widehat{A}x + \widehat{B} = \widehat{A}(x - \overline{x}) + \overline{Y},$$

gdje je $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, $\mathrm{E}[\overline{Y}] = a\overline{x} + b$, $\mathrm{V}[\overline{Y}] = \frac{1}{n}\sigma^2$. Iz (23), pak, slijedi da

$\overline{Y} \sim \mathrm{N}\left(a\overline{x}+b, \frac{1}{n}\sigma^2\right)$, a može se pokazati i da su $\widehat{A}$ i $\overline{Y}$ nezavisne slučajne varijable, zbog čega slučajnoj varijabli $\widehat{A}(x-\overline{x})+\overline{Y}$, kao linearnoj kombinaciji nezavisnih normalnih slučajnih varijabli, pripada normalna razdioba s očekivanjem

$$\mu_0 = a(x-\overline{x}) + a\overline{x} + b = ax + b$$

i varijancom

$$\sigma_0^2 = (x-\overline{x})^2 \frac{\sigma^2}{ns_x^2} + \frac{1}{n}\sigma^2 = \frac{\sigma^2}{n}\left[1+\left(\frac{x-\overline{x}}{s_x}\right)^2\right].$$

To znači da vrijedi

(32) $$\widehat{A}x + \widehat{B} \sim \mathrm{N}\left(ax+b, \frac{\sigma^2}{n}\left[1+\left(\frac{x-\overline{x}}{s_x}\right)^2\right]\right).$$

Nadalje, iz (26) i (32), pozivom na točku 7. iz V.6, proizlazi

(33) $$T_3 = \frac{\widehat{A}x+\widehat{B}-ax-b}{\frac{\sigma}{\sqrt{n}}\sqrt{1+\left(\frac{x-\overline{x}}{s_x}\right)^2}}\sqrt{\frac{n-2}{\frac{n-2}{\sigma^2}S^2}} = \frac{\widehat{A}x+\widehat{B}-ax-b}{S}\sqrt{\frac{n}{1+\left(\frac{x-\overline{x}}{s_x}\right)^2}} \sim \mathrm{t}(n-2),$$

a to omogućuje da se dobiju formule za rubove $g_1(x)$ i $g_2(x)$ intervala povjerenja pouzdanosti γ za nepoznatu vrijednost $\mu_t(x) = ax+b$ regresijske funkcije. Dobiva se

(34) $$\begin{cases} g_1(x) = \widehat{a}x + \widehat{b} - \tau_\gamma \frac{s}{\sqrt{n}}\sqrt{1+\left(\frac{x-\overline{x}}{s_x}\right)^2} \\ g_2(x) = \widehat{a}x + \widehat{b} + \tau_\gamma \frac{s}{\sqrt{n}}\sqrt{1+\left(\frac{x-\overline{x}}{s_x}\right)^2}. \end{cases}$$

Prema tome, u opisanom modelu, koji se zove *model jednodimenzionalne linearne regresije*, moguće je nepoznatu vrijednost regresijske funkcije $\mu_t(x) = ax+b$ nepristrano i konzistentno procijeniti pomoću vrijednosti $\widehat{a}x+\widehat{b}$ procjenitelja $\widehat{A}x+\widehat{B}$, pri čemu se može jamčiti s vjerojatnošću γ da apsolutna greška neće premašiti vrijednost

(35) $$d(x) = 2\tau_\gamma \frac{s}{\sqrt{n}}\sqrt{1+\left(\frac{x-\overline{x}}{s_x}\right)^2}, \; x \in \mathbf{R}.$$

Usporedbom formula (21) i (35) vidi se da su one vrlo slične. Umjesto koeficijenta z_γ u (21), koji se odnosi na standardnu normalnu razdiobu, u (35) stoji

koeficijent τ_γ, koji se odnosi na Studentovu razdiobu sa $n-2$ stupnja slobode. Umjesto nepoznatog parametra σ u (21), koji se u praksi zamjenjuje pripadnom procjenom s, u (35) stoji baš s.

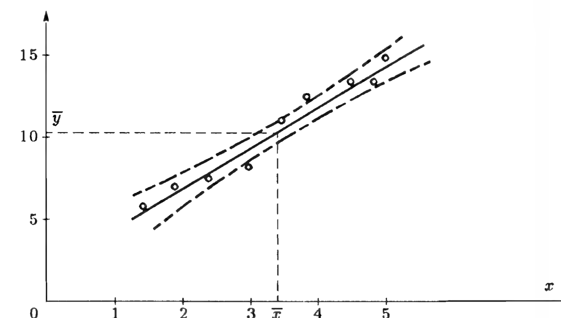
Napomenimo još jednom da formula (35) vrijedi za svako $n > 2$, dok formula (21) približno vrijedi za velike n.

2. primjer

Mjerenjem varijabli x i Y_x dobiveni su rezultati prikazani tablicom 1.

Tablica 1.

i	1	2	3	4	5	6	7	8	9
x_i	1,5	1,8	2,4	3,0	3,5	3,9	4,4	4,8	5,0
y_i	4,8	5,7	7,0	8,3	10,9	12,4	13,1	13,6	15,3



Slika 33. Prikaz regresijske ovisnosti za podatke iz tabl. 1.

Pogled na sl. 33. odmah nam sugerira da bi se mogao usvojiti model linearne regresije. Izvrše li se proračuni prema formulama (4), (5), (6), (16) i (17), dobiva se

$$\overline{x} = 3{,}37, \quad s_x = 1{,}21, \quad s_{xy} = 4{,}24,$$
$$\widehat{a} = 2{,}91, \quad \widehat{b} = 0{,}31, \quad s^2 = 0{,}47.$$

Prema tome, pravac regresije određen metodom najmanjih kvadrata ima jednadžbu $y = 2{,}91x + 0{,}31$.

Rubovi intervala povjerenja za parametar a, pouzdanosti $\gamma = 0{,}95$, na temelju formula (29), iznose

$$g_1 = 2{,}91 - 2{,}36 \cdot \frac{0{,}69}{1{,}21 \cdot \sqrt{9}} = 2{,}46\,, \quad g_2 = 3{,}36,$$

što znači da se, s vjerojatnošću od 95%, može jamčiti da se nepoznati koeficijent regresije a nalazi u intervalu $\langle 2{,}46\,; 3{,}36\rangle$.

Slično se izračunavaju rubovi intervala povjerenja pouzdanosti $\gamma = 0{,}95$ za parametar b. Na temelju formula (30) dobiva se $g_1 = -1{,}30$, $g_2 = 1{,}92$.

Formule (34) omogućuju da se izračuna interval povjerenja pouzdanosti $\gamma = 0{,}95$ za vrijednost $ax + b$ regresijske funkcije za svaki $x \in \mathbf{R}$. Tako se za $x = 2$ dobiva

$$g_1(2) = 2{,}91 \cdot 2 + 0{,}31 - 2{,}36\frac{0{,}69}{\sqrt{9}}\sqrt{1 + \left(\frac{2 - 3{,}37}{1{,}21}\right)^2} = 6{,}13 - 0{,}82 = 5{,}31,$$

$$g_2(2) = 6{,}13 + 0{,}82 = 6{,}95,$$

što znači da se, uz rizik od 5%, može smatrati da funkcija regresije u točki $x = 2$ ima vrijednost unutar intervala $\langle 5{,}31\,;\, 6{,}95\rangle$. Odgovarajuća širina intervala povjerenja je $d(2) = 1{,}64$. Za $x = \overline{x} = 3{,}37$ dobiva se najuži interval povjerenja $\langle 9{,}58\,;\, 10{,}66\rangle$, širine $d(\overline{x}) = d(3{,}37) = 1{,}08$.

Problem procjene nepoznatih parametara a, b i σ^2 u modelu jednodimenzionalne linearne regresije može se rješavati i metodom najveće vjerojatnosti (v. VI.3, formule (36) i (37)). Ovdje se, naime, može $\boldsymbol{t} = (a, b, \sigma^2)$ razmatrati kao nepoznati vektorski parametar koji varira po skupu

$$\boldsymbol{\Theta} = \{(a, b, \sigma^2) : a \in \mathbf{R}, b \in \mathbf{R}, \sigma > 0\},$$

dok se za funkciju vjerodostojnosti uzima

$$\boldsymbol{L}(\boldsymbol{t}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(ax_i + b - y_i)^2\right]. \tag{36}$$

Rješavanjem sustava jednadžbi

$$\frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial a} = 0\,, \qquad \frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial b} = 0\,, \qquad \frac{\partial \boldsymbol{L}(\boldsymbol{t})}{\partial \sigma^2} = 0\,,$$

po a, b i σ^2, dobiva se

$$a = \widehat{a} = \frac{s_{xy}}{s_x^2}\,, \qquad b = \widehat{b} = \overline{y} - \widehat{a}\overline{x}, \tag{37}$$

$$\sigma^2 = \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{a}x_i + \widehat{b} - y_i)^2 = \frac{n-2}{n}s^2. \tag{38}$$

Usporedbom (37) sa (4) i (5) vidi se da se procjene $\widehat{a}$ i $\widehat{b}$ nepoznatih parametara a i b, dobivene metodom najmanjih kvadrata i metodom najveće vjerojatnosti, poklapaju. Procjene s^2 i $\widehat{\sigma}^2$, za nepoznati parametar σ^2, razlikuju se samo u faktoru $\left(\frac{1}{n-2} \text{ umjesto } \frac{1}{n}\right)$, koji procjenitelj S^2 čini nepristranim, dok odgovarajući ML-procjenitelj $\widehat{\Sigma}^2$, s vrijednostima $\widehat{\sigma}^2$, nije nepristran.

Iz navedenoga je očigledno da je metoda najmanjih kvadrata općenitija od metode maksimalne vjerojatnosti, jer zahtijeva slabije pretpostavke (ne zahtijeva se normalna razdioba za greške $\mathcal{E}_i$). Osim toga MNK-procjenitelji, poput ML-procjenitelja, imaju mnoga dobra svojstva. Poznat je tzv. *Gauss-Markovljev teorem* (v. XII.3), koji izriče da u klasi svih linearnih nepristranih procjenitelja

za parametre a i b, MNK-procjenitelji imaju najmanje varijance. Iz (7) i (8) se, naime, vidi da su $\widehat{A}$ i $\widehat{B}$ linearne kombinacije slučajnih varijabli $Y_1, \ldots, Y_n$, tj. da pripadaju klasi linearnih procjenitelja, dok (11) i (12) pokazuju da je riječ o nepristranim procjeniteljima, pa Gauss-Markovljev teorem jamči da su MNK-procjenitelji $\widehat{A}$ i $\widehat{B}$ najefikasniji (v. VI.7) procjenitelji za parametre a i b, a također i da je $\widehat{A}x + \widehat{B}$ najefikasniji procjenitelj za $ax + b$ ($x \in \mathbf{R}$) u opisanoj klasi procjenitelja.

Time je još jače opravdana primjena metode najmanjih kvadrata u regresijskoj analizi, jer se pokazalo da su MNK-procjenitelji najefikasniji procjenitelji u vrlo opsežnoj klasi linearnih nepristranih procjenitelja.

Istaknimo još jedan važan praktični aspekt teorijskih rezultata dobivenih na temelju opisanih regresijskih modela. Određivanjem procjena $\widehat{a}$ i $\widehat{b}$ parametara a i b linearnog regresijskog modela moguće je, bar u načelu, procijeniti vrijednost regresijske funkcije $\mu_t(x) = ax + b$, za svaki $x \in \mathbf{R}$. Međutim, često u praktičnim zadacima postoje određena ograničenja na nezavisnu varijablu x (dopušteni su, recimo, samo pozitivni brojevi i sl.), tako da se postavlja zadatak procjene nepoznate vrijednosti $\mu_t(x)$ samo za $x \in A \subseteq \mathbf{R}$, gdje je A skup dopuštenih vrijednosti za iksove. Očigledno su $x_i \in A$ ($i = 1, \ldots, n$) i neka je $x_1 < x_2 < \ldots < x_n$. Ako je riječ o procjeni vrijednosti $\mu_t(x)$, za $x_1 \leq x \leq x_n$, onda se govori o *interpolaciji*, a ako je riječ o procjeni vrijednosti regresijske funkcije $\mu_t(x)$ za $x \in A$ i $x < x_1$, ili $x > x_n$, onda se govori o *ekstrapolaciji*.

Formule (20), odnosno (34), omogućuju da se uoči greška pri interpolaciji, odnosno ekstrapolaciji, iz čega se razabire da se točnost smanjuje udaljavanjem od središta $\overline{x}$ podataka o nezavisnoj varijabli. Već to pokazuje da je ekstrapolacija delikatniji problem od interpolacije. Poseban oprez u primjeni ekstrapolacije nužan je i zbog toga što je linearnost modela, donekle, očigledna za $x_1 \leq x \leq x_n$ (to jamči grafički prikaz podataka), ali za $x < x_1$ i $x > x_n$ to više nije očigledno iz grafičkog prikaza podataka, tako da se primjena linearne ekstrapolacije mora opravdati nekim drugim spoznajama o promatranoj pojavi, a ne samo izmjerenim podacima $(x_1, y_1), \ldots, (x_n, y_n)$.

Osim problema interpolacije i ekstrapolacije regresijske funkcije, može se postaviti i problem *prognoze* vrijednosti y_x slučajne varijable Y_x, za određenu vrijednost $x \in A$. Budući da je $Y_x = ax + b + \mathcal{E}_x$, čini se najlogičnijim nepoznate parametre a i b zamijeniti njihovim MNK-procjenama $\widehat{a}$ i $\widehat{b}$, a slučajnu varijablu $\mathcal{E}_x$ njenim očekivanjem $\mathrm{E}[\mathcal{E}_x] = 0$, tako da je $\widehat{y}_x = \widehat{a}x + \widehat{b}$ prognozirana vrijednost izlaza *(outputa)*, za ulaz *(input)* x. Odmah se postavlja i problem procjene greške prognoze. Prirodno je da se greška "mjeri" varijancom $\mathrm{V}[\widehat{A}x + \widehat{B} + \mathcal{E}_x]$. Iz pretpostavki regresijskog modela i formule (48) u V.6. proizlazi

$$V[\widehat{A}x+\widehat{B}+\mathcal{E}_x] = x^2V[\widehat{A}]+V[\widehat{B}]+V[\mathcal{E}_x]+2x\mathrm{Cov}(\widehat{A},\widehat{B})+2x\mathrm{Cov}(\widehat{A},\mathcal{E}_x)+2\mathrm{Cov}(\widehat{B},\mathcal{E}_x),$$

odnosno

$$V[\widehat{A}x + \widehat{B} + \mathcal{E}_x] = \frac{\sigma^2}{n}\left[3 + n + \left(\frac{x - \overline{x}}{s_x}\right)^2\right]. \tag{39}$$

Iako se nepoznata vrijednost regresijske funkcije $\mu_t(x)$ i nepoznata vrijednost y_x izlazne slučajne varijable Y_x procjenjuju istom veličinom $\widehat{y}_x = \widehat{a}x + \widehat{b}$, greške tih procjena su različite. Slično kao što su izvedene formule (34), mogu se, naime, izvesti (v. zad. 9) i formule za rubove g_1 i g_2 intervala povjerenja za nepoznatu vrijednost y_x. Uzimajući u obzir rezultat izražen formulom (39), dobiva se

$$(40)\quad g_1(x) = \hat{a}x + \hat{b} - \tau_\gamma \frac{s}{\sqrt{n}}\sqrt{3 + n + \left(\frac{x-\overline{x}}{s_x}\right)^2}$$
$$g_2(x) = \hat{a}x + \hat{b} + \tau_\gamma \frac{s}{\sqrt{n}}\sqrt{3 + n + \left(\frac{x-\overline{x}}{s_x}\right)^2},$$

iz čega se razabire da je pripadna širina intervala povjerenja

$$(41)\quad d(x) = 2\tau_\gamma \frac{s}{\sqrt{n}}\sqrt{3 + n + \left(\frac{x-\overline{x}}{s_x}\right)^2}.$$

Zanimljivo je primijetiti da procjena vrijednosti y_x, veličinom $\hat{y}_x = \hat{a}x + \hat{b}$, nije konzistentna u smislu definicije konzistentnosti iz VI.8, jer za $n \to \infty$ varijanca $V[\hat{A}x + \hat{B} + \mathcal{E}_x] \to \sigma^2$. Stoga se i događa da za velike n širina intervala povjerenja (41) ne teži nuli, već pozitivnoj veličini $2s\tau_\gamma$, za razliku od širine intervala povjerenja (35), koja teži nuli za $n \to \infty$.

3. Analiza rasipanja podataka

Regresijski modeli, posebno modeli linearne regresije, vrlo se mnogo primjenjuju u istraživanjima, opisivanjima i tumačenjima različitih praktičnih fenomena. Zato nije čudo da su se razvili raznoliki pristupi i interpretacije teorijskih modela radi jasnijeg, lakšeg i boljeg razumijevanja proučavanih fenomena. Sada će se opisati jedan takav pristup.

Polazeći od danog niza mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ i odgovarajuće procjene $\hat{\mu}_i = \hat{a}x_i + \hat{b}$ nepoznate vrijednosti regresijske funkcije $\mu_t(x_i) = ax_i + b$, definira se

$$(42)\quad \hat{\varepsilon}_i = y_i - \hat{\mu}_i, \quad i = 1, \ldots, n,$$

što se zove *reziduum* (residual). Veličina $\hat{\varepsilon}_i$, geometrijski interpretirana, pokazuje udaljenost između izmjerene vrijednosti y_i i prognozirane vrijednosti $\hat{\mu}_i = \hat{a}x_i + \hat{b}$ (v. sl. 31). To je, dakle, razlika između izmjerene vrijednosti izlazne varijable i one vrijednosti $\hat{\mu}_i$ koja se može objasniti funkcijskom vezom između ulazne i izlazne varijable. Dio $\hat{\varepsilon}_i$ izlazne varijable ($y_i = \hat{\mu}_i + \hat{\varepsilon}_i$) ne može se objasniti funkcijskom ovisnošću izlaza o ulazu, već potječe od djelovanja slučajnih faktora (slučajne greške $\mathcal{E}_i$).

Imajući na umu (16) i (38), odmah se vidi da se može pisati

$$(43)\quad s_y^2 = \hat{\sigma}^2 + \frac{s_{xy}^2}{s_x^2}.$$

Veličina s_y^2 opisuje rasipanje izlaznih podataka (ipsilona) oko njihove sredine $\overline{y}$, dok $\hat{\sigma}^2$ opisuje, kao što se vidi iz (38), rasipanje izlaznih podataka oko procijenjene regresijske funkcije. Uz nešto složeniji račun (v. zad. 6), dokazuje se da je

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{a}x_i + \hat{b} - \overline{y})^2 = \frac{s_{xy}^2}{s_x^2},$$

pa se vidi da se veličina

$$(44)\quad \hat{\sigma}_0^2 = \frac{s_{xy}^2}{s_x^2}$$

može interpretirati kao mjera rasipanja vrijednosti procijenjene regresijske funkcije (prognoziranih vrijednosti izlaza) oko $\overline{y}$. U tom se svjetlu relacija (43), zapisana kao

$$(45)\quad s_y^2 = \hat{\sigma}^2 + \hat{\sigma}_0^2,$$

može interpretirati tako da se kaže da je rasipanje izlaznih podataka oko njihove aritmetičke sredine jednako zbroju rasipanja uzrokovanog regresijskom ovisnošću (funkcijom $x_i \mapsto \hat{\mu}_i$) i rasipanja uzrokovanog slučajnom greškom, tzv. *rezidualnog rasipanja*. Veličina

$$(46)\quad R^2 = \frac{\hat{\sigma}_0^2}{s_y^2} = 1 - \frac{\hat{\sigma}^2}{s_y^2}$$

zove se *koeficijent determinacije*. Očigledno je

$$(47)\quad 0 \leq R^2 \leq 1.$$

Za podatke iz 2. primjera nalazi se da je

$$s_y^2 = 12{,}72, \quad \hat{\sigma}^2 = 0{,}37, \quad \hat{\sigma}_0^2 = 12{,}35,$$

pa se iz (46) odmah dobiva odgovarajući koeficijent determinancije $R^2 = 0{,}97$, što bismo protumačili tako da 97% rasipanja izlaznih podataka potječe od funkcijske ovisnosti $x \mapsto 2{,}91x + 0{,}31$, a samo 3% otpada na rezidualno rasipanje, koje se ponegdje zove i *neobjašnjeno rasipanje*. Preveliko neobjašnjeno rasipanje obično upućuje istraživača na promjenu regresijskog modela, iako ono može biti i posljedica slabe koreliranosti između ulaznih i izlaznih podataka.

Formula (46) može se, naime, zapisati i u obliku

$$(48)\quad R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2},$$

iz čega se, usporedbom s formulom (27) iz III.4, razabire da je koeficijent determinacije isto što i kvadrat koeficijenta korelacije za podatke $(x_1, y_1), \ldots, (x_n, y_n)$.

Sada se možemo pitati uz koje uvjete se postižu krajnje vrijednosti 0 i 1 koeficijenta determinacije. Iz (46) se vidi da će $R^2 = 0$ biti za $\hat{\sigma}_0^2 = 0$, ili za $\hat{\sigma}^2 = s_y^2$. Jedan i drugi uvjet impliciraju $s_{xy} = 0$, a to znači da su x_i i y_i ($i = 1, \ldots, n$) nekorelirani podaci (v. III.4. i III.5). Regresijski pravac u tom je slučaju usporedan s apscisnom osi ($\hat{a} = 0$), što upućuje na zaključak da ulazna (nezavisna) varijabla x ne utječe na izlaznu slučajnu varijablu Y. Ukupno rasipanje izlaznih podataka je, zapravo, rezidualno rasipanje, jer se ništa od njega ne može objasniti funkcijskom vezom između x i Y. Poznavanje ulazne vrijednosti x ne omogućuje da se bilo što novo kaže o pripadnom izlazu y_x, što već ne bi bilo moguće reći i bez poznavanja x.

Drugi krajnji slučaj $R^2 = 1$ postiže se za $\widehat{\sigma}^2 = 0$. To znači da nema rezidualnog rasipanja i da ukupno rasipanje izlaznih podataka potječe od funkcijske ovisnosti između iksova (x_i) i ipsilona (y_i) oblika $y_i = \widehat{a}x_i + \widehat{b}$ $(i = 1, \ldots, n)$, pri čemu je $\widehat{a}^2 = \frac{s_y^2}{s_x^2} \neq 0$. Budući da je $s^2 = \frac{n}{n-2}\widehat{\sigma}^2$, u tom je slučaju i $s^2 = 0$, pa se iz (40) razabire da je tada $g_1(x) = g_2(x) = \widehat{a}x + \widehat{b}$ i $d(x) = 0$, što pokazuje da se s proizvoljnom pouzdanošću, recimo $\gamma = 0{,}99$, može prognozirati izlazna vrijednost $\widehat{y}_x = \widehat{a}x + \widehat{b}$.

U statističkoj literaturi, posebno onoj u kojoj se uglavnom obrađuje primjena teorije statističkog zaključivanja, uobičajen je određeni način tabličnog prikazivanja analize rasipanja podataka.

Tablica 2.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Srednje kvadratno odstupanje	Koeficijent determinacije
regresijska funkcija (model)	1	$\sum_{i=1}^{n}(\widehat{\mu}_i - \overline{y})^2$	$\widehat{\sigma}_0^2$	
rezidualno rasipanje (slučajna greška)	$n-2$	$\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$	$\widehat{\sigma}^2$	
ukupno rasipanje	$n-1$	$\sum_{i=1}^{n}(y_i - \overline{y})^2$	s_y^2	$R^2 = \frac{\widehat{\sigma}_0^2}{s_y^2}$

Tablica 2. obično se zove ***tablica analize varijance*** u jednodimenzionalnome linearnom regresijskom modelu. Tablica analize varijance redovito se, inače, primjenjuje u modelima analize varijance o kojima će biti riječi u nastavku.

4. Testiranje hipoteza o koeficijentu regresije

U 1. i 2. primjeru imali smo podatke koji su očigledno pokazivali da je riječ o linearnoj regresiji s regresijskim koeficijentom značajno različitim od nule, tako da je promjena nezavisne varijable prouzročila, preko regresijske funkcije, odgovarajuću promjenu izlazne slučajne varijable. Ako se, međutim, dobije mala apsolutna vrijednost procjene $\widehat{a}$ za nepoznati koeficijent regresije a, može se posumnjati da je stvarna vrijednost koeficijenta regresije zapravo nula, što bi značilo da regresijski model ima oblik

(49) $$Y_i = b + \mathcal{E}_i, \quad i = 1, \ldots, n,$$

tj. da je izmjerena izlazna vrijednost y_i zapravo neovisna o ulaznoj vrijednosti x_i. Može se tada pisati

$$y_i = b + \varepsilon_i, \quad i = 1, \ldots, n,$$

i reći da izlaz y_i nastaje kao zbroj konstante b i vrijednosti ε_i slučajne greške $\mathcal{E}_i \sim \mathrm{N}(0, \sigma^2)$. U tom je slučaju

$$\mathrm{E}[Y_i] = b, \quad \mathrm{V}[Y_i] = \mathrm{V}[\mathcal{E}_i] = \sigma^2,$$

iz čega se vidi da je regresijska linija pravac $y = b$, pa se problem regresijske analize u tom slučaju svodi na problem procjene nepoznatog očekivanja i nepoznate varijance slučajne varijable $Y \sim \mathrm{N}(b, \sigma^2)$, što je inače detaljno obrađeno u VI.4.

Ostalo je, dakle, da se riješi problem testiranja hipoteze $\mathrm{H}_0\colon a = 0$, prema nekoj od alternativnih hipoteza ($\mathrm{H}_1\colon a \neq 0$, ili $\mathrm{H}_1\colon a < 0$, ili $\mathrm{H}_1\colon a > 0$). Budući da se problem bitno ne komplicira, rješavat će se zadatak određivanja test-statistike i kritičnog područja razine značajnosti α, pri testiranju hipoteze $\mathrm{H}_0\colon a = a_0$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a \neq a_0$, gdje je a_0 zadani realan broj. Pretpostavlja se, dakako, da imamo niz od n $(n > 2)$ mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$, na što se može primijeniti model jednodimenzionalne linearne regresije. To omogućuje da se iskoristi činjenica izražena formulom (27), što znači da u uvjetima istinitosti hipoteze H_0 vrijedi

(50) $$T_1 = \frac{\widehat{A} - a_0}{S} s_x \sqrt{n} \sim \mathrm{t}(n-2).$$

Slučajna varijabla T_1 iz (50) uzet će se kao test-statistika, što se može opravdati činjenicom da vrijednost

(51) $$t_1 = \frac{\widehat{a} - a_0}{s} s_x \sqrt{n}$$

upućuje na odstupanje procjene $\widehat{a}$, dobivene na temelju danih podataka, od pretpostavljene vrijednosti a_0 koeficijenta regresije a.

Odmah se zaključuje da će kritično područje razine značajnosti α biti određeno uvjetom

(52) $$|t_1| \geq c_0,$$

pri čemu je c_0 $(c_0 > 0)$ određeno tako da u Studentovoj razdiobi sa $n-2$ stupnja slobode vrijedi

$$\mathrm{P}(|T_1| \geq c_0) = \alpha,$$

odnosno

(53) $$c_0 = G_{n-2}^{-1}\left(1 - \frac{\alpha}{2}\right).$$

U 1. primjeru imali smo $\widehat{a} = 0{,}6$, što bi nas eventualno moglo navesti na pomisao da testiramo hipotezu $\mathrm{H}_0\colon a = 0$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a \neq 0$, uz razinu značajnosti $\alpha = 0{,}95$. Praktički bi to značilo da sumnjamo u postavku da se s porastom dobi povisuje krvni tlak, te postavljamo hipotezu da dob ne utječe na krvni tlak.

Iz činjenice da je $n = 100$, $s_x = 12$ i $s = 5{,}7$ dobiva se $t_1 = \frac{0{,}6 - 0}{5{,}7} \cdot 12 \cdot \sqrt{100} = 12{,}63$. U tabl. V. u Dodatku ne može se očitati $G_{98}^{-1}(0{,}975)$, ali se zna da se, za $n > 30$, Studentova razdioba $\mathrm{t}(n)$ može aproksimirati standardnom normalnom razdiobom $\mathrm{N}(0, 1)$, tako da je $G_{98}^{-1}(0{,}975) \approx \Phi^{-1}(0{,}975) = 1{,}96$, što pokazuje da je kritično područje $\langle -\infty; -1{,}96] \cup [1{,}96; \infty\rangle$. Očigledno je da dobivena vrijednost $t_1 = 12{,}63$, test-statistike T_1, duboko upada u kritično područje i zato hipotezu H_0

treba odbaciti. Reklo bi se da navedeni podaci o krvnom tlaku osoba nikako ne opravdavaju hipotezu da krvni tlak (Y) ne ovisi o dobi (x).

U 2. primjeru dobili smo regresijski pravac s procjenom $\hat{a} = 2{,}91$ za koeficijent regresije a, pa bismo mogli testirati hipotezu $\mathrm{H}_0\colon a = 3$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a \neq 3$, uz razinu značajnosti $\alpha = 0{,}05$. Sada je $n = 9$, $s_x = 1{,}21$ i $s = 0{,}69$, pa test-statistika T_1 poprima vrijednost $t_1 = \dfrac{2{,}91 - 3}{0{,}69} \cdot 1{,}21\sqrt{9} = -0{,}47$. Kritično područje testa određeno je vrijednošću $c_0 = G_7^{-1}(0{,}975) = 2{,}365$ (v. tabl. V. u Dodatku), a budući da je $|t_1| = 0{,}47 < 2{,}365$, vidi se da vrijednost test-statistike ne pada u kritično područje, što upućuje na zaključak da hipotezu H_0 ne treba odbaciti.

Problem testiranja hipoteze o koeficijentu regresije može se, dakako, postaviti i kao problem jednorubnog testa, gdje se testira hipoteza $\mathrm{H}_0\colon a = a_0$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a > a_0$. U tom je slučaju kritično područje razine značajnosti α određeno uvjetom

(54) $$t_1 \geq G_{n-2}^{-1}(1 - \alpha).$$

Uzme li se kao alternativna hipoteza $\mathrm{H}_1\colon a < a_0$, kritično područje određeno je uvjetom

(55) $$t_1 \leq G_{n-2}^{-1}(\alpha).$$

Tako, na primjer, testiramo li na podacima iz 1. primjera hipotezu $\mathrm{H}_0\colon a = 0$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a > 0$, uz razinu značajnosti $\alpha = 0{,}05$, kritično je područje određeno uvjetom $t_1 \geq G_{98}^{-1}(0{,}95) \approx \Phi^{-1}(0{,}95) = 1{,}65$. Budući da je $t_1 = 12{,}63$, očigledno je da treba odbaciti H_0 i prihvatiti H_1, tj. zaključiti da su krvni tlak i dob povezani linearnom regresijom s pozitivnim koeficijentom regresije.

Upravo opisani postupak testiranja različitih hipoteza o koeficijentu regresije zasniva se na test-statistici $T_1 \sim \mathrm{t}(n-2)$ i njime je omogućeno provjeriti nul-hipotezu $\mathrm{H}_0 : a = a_0$, za proizvoljno $a_0 \in \mathbf{R}$. Sada ćemo opisati još jedan test kojim se, doduše, može testirati samo nul-hipoteza oblika $\mathrm{H}_0\colon a = 0$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a \neq 0$.

Test se zasniva na test-statistici

(56) $$V = (n-2)\frac{\widehat{\Sigma}_0^2}{\widehat{\Sigma}^2} = n s_x^2 \frac{\widehat{A}^2}{S^2},$$

gdje je $\widehat{\Sigma}_0^2$ statistika s vrijednostima $\widehat{\sigma}_0^2$, a $\widehat{\Sigma}^2$ statistika s vrijednostima $\widehat{\sigma}^2$. U uvjetima istinitosti hipoteze H_0 slučajnoj varijabli V pripada F-razdioba sa $(1, n-2)$ stupnjeva slobode. Može se, dakle, pisati

(57) $$V \sim \mathrm{F}(1, n-2).$$

Da bi se obrazložilo (56) i (57) razmišljat će se ovako: Slučajna varijabla V kvocijent je slučajnih varijabli $\widehat{\Sigma}_0^2$ i $\widehat{\Sigma}^2$, odnosno $\widehat{A}^2$ i S^2, za koje se vidi da su nastale kao sume kvadrata normalnih slučajnih varijabli, pa se, u skladu s rezultatima navedenim u točkama 5. i 8. iz V.6, može dokazati da je V slučajna varijabla F-razdiobe, što je izrečeno relacijom (57).

Ako je hipoteza $\mathrm{H}_0\colon a = 0$ stvarno istinita, onda se može očekivati da će i vrijednost $\widehat{a}$ procjenitelja $\widehat{A}$ za nepoznati koeficijent regresije a biti vrlo blizu nule,

pa će i vrijednost

(58) $$v = (n-2)\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}^2} = n s_x^2 \frac{\widehat{a}^2}{s^2}$$

test-statistike V iz (56) i (57) biti blizu nule.

Dobije li se na danim podacima $(x_1, y_1), \ldots, (x_n, y_n)$ prevelika vrijednost v test-statistike V, hipoteza H_0 će se odbaciti. Prema tome, kritično područje razine značajnosti α bit će određeno uvjetom $v \geq c_0$, gdje je c_0 određeno tako da vrijedi

$$\mathrm{P}(V \geq c_0) = \alpha.$$

Koristeći se tabl. VII. iz Dodatka, može se, u konkretnom slučaju, odrediti c_0 primjenom formule

(59) $$c_0 = F_{1,n-2}^{-1}(1 - \alpha),$$

gdje je $F_{1,n-2}^{-1}$ inverzna funkcija od f.r.v. za F-razdiobu sa $(1, n-2)$ stupnjeva slobode.

Zanimljivo je da iz (51) i (58) slijedi da je $t_1^2 = v$ pri nul-hipotezi $\mathrm{H}_0\colon a = 0$, tj. kvadrat vrijednosti test-statistike T_1 iz (50), kojoj pripada Studentova razdioba $\mathrm{t}(n-2)$, jednak je vrijednosti test-statistike V iz (56), kojoj pripada F-razdioba $\mathrm{F}(1, n-2)$. To je, zapravo, posljedica općeg stavka izraženog relacijom (v. zad. 14)

(60) $$T \sim \mathrm{t}(n) \Rightarrow V = T^2 \sim \mathrm{F}(1, n), \quad n > 2.$$

Sada jednostavno možemo, na podacima iz 1. primjera testirati hipotezu $\mathrm{H}_0 : a = 0$, prema alternativnoj hipotezi $\mathrm{H}_1\colon a \neq 0$, pomoću test-statistike V. Maloprije smo izračunali $t_1 = 12{,}63$, pa odmah možemo zaključiti da je $v = t_1^2 \approx 160$. Za $\alpha = 0{,}05$ iz tabl. VII. u Dodatku dobiva se $c_0 = F_{1,98}^{-1}(0{,}95) \approx 6{,}96$, iz čega se vidi da je $t_1 > c_0$, što znači da vrijednost test-statistike i sada duboko upada u kritično područje $[6{,}96; \infty\rangle$, pa hipotezu H_0 treba odbaciti.

Zadaci

1. Izvedite rješenja, izražena formulama (4) i (5), kojima su definirane procjene za nepoznate parametre a i b u smislu metode najmanjih kvadrata.
2. Dokažite formule (11), (12) i (13).
3. Izvedite formule (20) za granice intervala povjerenja pouzdanosti γ, za nepoznatu vrijednost $ax + b$ regresijske funkcije.
4. Izvedite formule (29), (30) i (34).
5. Dokažite da su formulama (37) i (38) izražene ML-procjene za nepoznate parametre a, b i σ^2 jednodimenzionalnoga linearnog regresijskog modela.
6. Dokažite da za veličine, definirane u XI.1. i XI.3. vrijedi:

   a) $\widehat{\varepsilon}_i = y_i - \overline{y} - \widehat{a}(x_i - \overline{x})$,

   b) $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$,

c) $\displaystyle\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 = \widehat{\sigma}^2 = s_y^2 - \widehat{a}^2 s_x^2 = s_y^2 - \left(\frac{s_{xy}}{s_x}\right)^2$,

d) $\displaystyle\frac{1}{n}\sum_{i=1}^{n}(\widehat{a}x_i + \widehat{b} - \overline{y})^2 = \widehat{a}^2 s_x^2 = \frac{s_{xy}^2}{s_x^2}$.

7. Dokažite da u jednodimenzionalnome linearnom regresijskom modelu vrijedi

a) $\displaystyle \mathrm{Cov}(\widehat{A}, Y_i) = \mathrm{Cov}(\widehat{A}, \mathcal{E}_i) = \frac{\sigma^2}{ns_x^2}(x_i - \overline{x})$,

b) $\displaystyle \mathrm{Cov}(\widehat{B}, Y_i) = \mathrm{Cov}(\widehat{B}, \mathcal{E}_i) = \frac{\sigma^2}{n}\left[1 - \frac{\overline{x}}{s_x^2}(x_i - \overline{x})\right]$,

c) ako je $\displaystyle \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, onda je $\mathrm{Cov}(\widehat{A}, \overline{Y}) = 0$ i $\displaystyle \mathrm{Cov}(\widehat{B}, \overline{Y}) = \frac{\sigma^2}{n}$.

8. Neka su $\widehat{A}_1 = \sum_{i=1}^{n} \alpha_i Y_i$ i $\widehat{B}_1 = \sum_{i=1}^{n} \beta_i Y_i$ $(\alpha_i, \beta_i \in \mathbf{R}, i = 1, \dots, n)$ *nepristrani linearni procjenitelji (NL-procjenitelji)* za nepoznate parametre a i b jednodimenzionalnoga linearnog regresijskog modela.

a) Dokažite da se uvjet nepristranosti procjenitelja $\widehat{A}_1$ i $\widehat{B}_1$ može zapisati u obliku

$$\sum_{i=1}^{n}\alpha_i x_i = 1, \quad \sum_{i=1}^{n}\alpha_i = 0, \quad \sum_{i=1}^{n}\beta_i x_i = 0 \quad \text{i} \quad \sum_{i=1}^{n}\beta_i = 1.$$

b) Dokažite da zahtjev za minimalnost varijance procjenitelja $\widehat{A}_1$ i $\widehat{B}_1$ implicira

$$\alpha_i = \frac{(x_i - \overline{x})}{ns_x^2}, \quad \beta_i = \frac{1}{n}\left[1 - \frac{\overline{x}}{s_x^2}(x_i - \overline{x})\right].$$

9. Izvedite formule (40).

10. Dokažite da se test-statistika T_1 iz (50), za testiranje hipoteze $\mathrm{H}_0\!: a = a_0$, može izvesti metodom primjene intervala povjerenja opisanom u VIII.7.

11. Načinite tablicu analize varijance za podatke iz:

a) 1. primjera, b) 2. primjera.

12. U priloženoj tablici x_i označuje godišnji dohodak (u dolarima) po stanovniku, a y_i postotak nepismenih među odraslim stanovništvom u određenoj afričkoj državi.

x_i	110	370	380	500	500
y_i	85	75	73	63	61

a) Prikažite dane podatke točkama u pravokutnom koordinatnom sustavu.

b) Nađite procjene $\widehat{a}$, $\widehat{b}$ i s^2 za odgovarajuće parametre linearnoga regresijskog modela.

c) Prognozirajte postotak nepismenih na razini nacionalnog dohotka od 200 dolara i od 1 000 dolara. Nađite pripadne varijance za te procjene.

d) Odredite granice intervala povjerenja pouzdanosti 95% za nepoznati koeficijent regresije a.

e) Ucrtajte u koordinatni sustav procijenjeni regresijski pravac i krivulje koje opisuju odgovarajući 90-postotni interval povjerenja.

f) Načinite odgovarajuću tablicu analize varijance.

g) Testirajte, uz razinu značajnosti 5%, hipotezu da nacionalni dohodak ne utječe na postotak nepismenih.

13. Podaci o ulaznoj i izlaznoj varijabli zadani su ovom tablicom

x_i	159	160	161	162	166	168	172
y_i	24	23	24	23	23	24	24
	24	25	24	24	23	29	25
		27	25	24	25	30	
			26	29		31	

a) Nađite procjene $\widehat{a}$, $\widehat{b}$ i s^2 za parametre linearnoga regresijskog modela.

b) Načinite odgovarajuću tablicu analize varijance.

c) Testirajte hipotezu $\mathrm{H}_0\!: a = 0$, prema alternativnoj hipotezi $\mathrm{H}_1\!: a \neq 0$, primjenom test-statistika T_1 i V, uz razinu značajnosti $\alpha = 0{,}05$.

14. Dokažite relaciju (60). Uputa: Iskoristite rezultate točke 5, 7. i 8. iz V.6.

Važniji rezultati regresijske analize

Model	$Y_i = ax_i + b + \mathcal{E}_i\,, \quad i = 1,\dots,n$
Standardne pretpostavke	1. x_i su vrijednosti neslučajne varijable 2. $\mathrm{E}[\mathcal{E}_i] = 0$ 3. $\mathrm{V}[\mathcal{E}_i] = \sigma^2$ 4. $\mathrm{Cov}(\mathcal{E}_i, \mathcal{E}_j) = 0, \quad \text{za } i \neq j$
Vrijednosti MNK-procjenitelja	$\widehat{a} = \dfrac{s_{xy}}{s_x^2}\,, \quad \widehat{b} = \overline{y} - \widehat{a}\overline{x}\,, \quad s^2 = \dfrac{n}{n-2}\left(s_y^2 - \dfrac{s_{xy}^2}{s_x^2}\right)$
Svojstva MNK-procjenitelja	$\mathrm{E}[\widehat{A}] = a, \quad \mathrm{E}[\widehat{B}] = b, \quad \mathrm{E}[S^2] = \sigma^2$ $\mathrm{V}[\widehat{A}] = \dfrac{\sigma^2}{n\,s_x^2}\,, \quad \mathrm{V}[\widehat{B}] = \dfrac{\sigma^2}{n}\left(1 + \dfrac{\overline{x}^2}{s_x^2}\right)$ $\mathrm{E}[\widehat{A}x + \widehat{B}] = ax + b\,, \quad \mathrm{V}[\widehat{A}x + \widehat{B}] = \dfrac{\sigma^2}{n}\left[1 + \dfrac{(x-\overline{x})^2}{s_x^2}\right]$
Koeficijent determinacije	$R^2 = \dfrac{s_{xy}^2}{s_x^2\,s_y^2}\,, \quad 0 \leq R^2 \leq 1$
Dodatna pretpostavka	$\mathcal{E}_i \sim \mathrm{N}(0, \sigma^2)$
Posljedice dodatne pretpostavke	$\widehat{A} \sim \mathrm{N}\left(a, \dfrac{\sigma^2}{n\,s_x^2}\right)\,, \quad \widehat{B} \sim \mathrm{N}\left(b, \dfrac{\sigma^2}{n}\left(1 + \dfrac{\overline{x}^2}{s_x^2}\right)\right)$ $\dfrac{n-2}{\sigma^2}\,S^2 \sim \chi^2(n-2)$ rubovi intervala povjerenja pouzdanosti γ za regresijski koeficijent a su $g_{1,2} = \widehat{a} \mp \tau_\gamma \dfrac{s}{s_x\sqrt{n}}$ rubovi intervala povjerenja pouzdanosti γ za nepoznatu vrijednost $ax + b$ su $g_{1,2} = \widehat{a}x + \widehat{b} \mp \tau_\gamma \dfrac{s}{\sqrt{n}}\sqrt{1 + \left(\dfrac{x-\overline{x}}{s_x}\right)^2}$

XII. Višestruka regresija

1. Model višedimenzionalne regresije

U svim regresijskim modelima razmotrenim u XI. poglavlju pretpostavljalo se da izlazna slučajna varijabla Y ovisi o jednoj ulaznoj (nezavisnoj) varijabli x, pa se prirodno nameće zadatak da se izgrade i istraže i modeli u kojima će izlaz Y ovisiti o više ulaznih varijabli $x^{(1)},\dots,x^{(r)}$ $(r \in \mathbb{N})$.

Prije nego što apstraktno i precizno definiramo model ***višedimenzionalne*** ili ***višestruke regresije*** (*multiple regression*), formulirat ćemo jedan konkretan primjer koji će poslužiti da se lakše i bolje shvate apstraktni pojmovi u vezi s tim modelom.

1. primjer

Osnovne sirovine za proizvodnju betona su cement, agregat (pijesak ili šljunak) i voda, pa o njima uglavnom ovisi tlačna čvrstoća, mjerena na betonskim kockama određene dimenzije, izrađenim po propisanim postupcima. Da bi se istražila ovisnost tlačne čvrstoće Y o upotrijebljenim sirovinama ($x^{(1)}$ – količina cementa u kilogramima, $x^{(2)}$ – količina agregata u kilogramima i $x^{(3)}$ – vodocementni faktor koji pokazuje omjer cementa i vode), načinjeno je $n = 7$ betonskih kocki od različitih mješavina sirovina i na svakoj je izmjerena vrijednost $y_i (i = 1,\dots,7)$ tlačne čvrstoće u megapaskalima (MPa). Rezultati su prikazani u tabl. 1

Tablica 1.

$x_i^{(1)}$	$x_i^{(2)}$	$x_i^{(3)}$	y_i
200	2123	0,600	29,0
220	2090	0,560	30,5
250	2060	0,510	33,8
280	2014	0.470	37,5
300	1993	0,450	39,8
320	1967	0,430	41,3
350	1926	0,405	42,3

Poznata je činjenica da je tlačna čvrstoća Y, u slučaju da se beton proizvodi po istoj "recepturi" (s ustaljenim udjelom pojedinih sirovina), slučajna varijabla za koju se obično pretpostavlja normalna razdioba. Zato ćemo reći da su izlazne vrijednosti y_i posljedica djelovanja određene funkcijske ovisnosti tlačne čvrstoće Y o količinama $x^{(1)}$, $x^{(2)}$ i $x^{(3)}$ sirovina i o slučajnoj komponenti (greški) $\mathcal{E}$, koja u sebi uključuje brojne druge faktore koji također utječu na tlačnu čvrstoću, i za koju

vjerujemo da se podvrgava određenoj statističkoj zakonitosti. Jednadžba

$$Y = \mu_t(x^{(1)}, x^{(2)}, x^{(3)}) + \mathcal{E} \tag{1}$$

usvaja se kao matematički model za opisivanje karaktera ovisnosti između tlačne čvrstoće betona i količina cementa, agregata i vode. Funkcija

$$(x^{(1)}, x^{(2)}, x^{(3)}) \mapsto \mu_t(x^{(1)}, x^{(2)}, x^{(3)})$$

realna je funkcija triju realnih varijabli, određena vrijednošću parametra $\boldsymbol{t} \in \boldsymbol{\Theta}$ ($\boldsymbol{\Theta}$ je skup dopuštenih vrijednosti za parametar $\boldsymbol{t}$ koji je redovito vektorski parametar).

Uzme li se, na primjer

$$\mu_{\boldsymbol{t}}(x^{(1)}, x^{(2)}, x^{(3)}) = a_1 x^{(1)} + a_2 x^{(2)} + a_3 x^{(3)} + a_4 ,$$

bit će $\boldsymbol{t} = (a_1, a_2, a_3, a_4) \in \mathbf{R}^4$, pa se prirodno nameće problem da se, na temelju podataka iz tabl. 1, odredi procjena $\widehat{\boldsymbol{t}} = (\widehat{a}_1, \widehat{a}_2, \widehat{a}_3, \widehat{a}_4)$ za nepoznati vektorski parametar $\boldsymbol{t}$, kao i procjene za odgovarajuće parametre slučajne varijable $\mathcal{E}$. Time se dobiva konkretni matematički odnos koji omogućuje da se uoči priroda ovisnosti tlačne čvrstoće betona o udjelu pojedinih sirovina, što je i postavljeno kao glavni cilj istraživanja.

Radi konciznijeg i jasnijeg formuliranja općeg modela višedimenzionalne regresije prikladno je uvesti vektorske i matrične oznake. Neka je, dakle, $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(r)})$ ($r \in \mathbf{N}$) r-dimenzionalna ulazna (neslučajna) vektorska varijabla, pa će $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(r)})$ označivati i-tu ($i = 1, \ldots, n$) vrijednost ulazne vektorske varijable, za koju je y_i pripadna vrijednost izlazne slučajne varijable Y_i. Neka je, nadalje, $\boldsymbol{x} \mapsto \mu_{\boldsymbol{t}}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbf{R}^r$, zadana funkcija određena vektorskim parametrom $\boldsymbol{t} \in \mathbf{R}^s$ ($s \in \mathbf{N}$) i neka su $\mathcal{E}_i$ ($i = 1, \ldots, n$) nezavisne slučajne varijable s očekivanjem $E[\mathcal{E}_i] = 0$ i varijancom $V[\mathcal{E}_i] = \sigma^2 > 0$, onda je, slično kao u XI.1, relacijom

$$Y_i = \mu_{\boldsymbol{t}}(\boldsymbol{x}_i) + \mathcal{E}_i , \qquad i = 1, \ldots, n , \tag{2}$$

definiran ***r-dimenzionalni regresijski model*** s regresijskom funkcijom $\boldsymbol{x} \mapsto \mu_{\boldsymbol{t}}(\boldsymbol{x})$.

I sada se može reći da je regresijskim modelom (2) opisana ona realna situacija gdje se smatra da je izmjerena izlazna vrijednost y_i nastala zbog postojanja funkcijske ovisnosti o ulaznoj vektorskoj varijabli $\boldsymbol{x}$ uz dodatak slučajne greške $\mathcal{E}_i$.

Jednadžbom $y = \mu_{\boldsymbol{t}}(\boldsymbol{x})$, općenito je određena neka apstraktna ploha u $(r+1)$-dimenzionalnom prostoru $\mathbf{R}^{r+1}$, koja se zove *regresijska ploha*.

Glavni je problem regresijske analize i u ovom slučaju, kao i u jednodimenzionalnom slučaju, nalaženje dobrih procjenitelja za nepoznate parametre $\boldsymbol{t}$ i σ^2, čije će se vrijednosti $\widehat{\boldsymbol{t}}$ i $\widehat{\sigma}^2$ računati na temelju danog niza podataka $(\boldsymbol{x}_i, y_i)$ ($i = 1, \ldots, n$). U ovoj situaciji slučajnim uzorkom smatra niz $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_n, Y_n)$, gdje su $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ vrijednosti ulazne (neslučajne) vektorske varijable, a $Y_1, \ldots, Y_n$ međusobno nezavisne slučajne varijable za koje vrijedi

$$E[Y_i] = \mu_{\boldsymbol{t}}(x_i) , \qquad V[Y_i] = \sigma^2 , \qquad i = 1, \ldots, n . \tag{3}$$

Za određivanje procjene $\widehat{\boldsymbol{t}}$ nepoznatog parametra $\boldsymbol{t}$ regresijske funkcije $\mu_{\boldsymbol{t}}$, primijenit će se i ovaj puta metoda najmanjih kvadrata. Treba, dakle, $\widehat{\boldsymbol{t}}$ odrediti tako da vrijedi

$$\min_{\boldsymbol{t} \in \boldsymbol{\Theta}} \sum_{i=1}^{n} [y_i - \mu_{\boldsymbol{t}}(\boldsymbol{x}_i)]^2 = \sum_{i=1}^{n} [y_i - \mu_{\widehat{\boldsymbol{t}}}(\boldsymbol{x}_i)]^2 . \tag{4}$$

Izražavajući se jezikom geometrije, svako mjerenje $(\boldsymbol{x}_i, y_i)$ može se interpretirati kao točka u $(r+1)$-dimenzionalnom prostoru $\mathbf{R}^{r+1}$, pa se određivanje procjene $\widehat{\boldsymbol{t}}$, u smislu metode najmanjih kvadrata, može shvatiti kao pronalaženje one plohe $y = \mu_{\widehat{\boldsymbol{t}}}(\boldsymbol{x})$, u skupu svih dopuštenih regresijskih ploha, koja se najbolje prilagođuje izmjerenim podacima $(\boldsymbol{x}_i, y_i)$ ($i = 1, \ldots, n$), tj. najbliža je odgovarajućim točkama prostora $\mathbf{R}^{r+1}$ u smislu relacije (4). (Zbroj kvadrata udaljenosti točaka od te plohe je minimalan.)

Odmah se, naravno, postavlja pitanje da li tako formulirani problem uopće ima rješenje, odnosno uz koje dodatne pretpostavke postoji jednoznačno rješenje problema. U prvom redu nužno je postaviti pretpostavke o regresijskoj funkciji, kojima se specificira tip funkcijske zavisnosti, iz čega će proizaći i dimenzija parametra $\boldsymbol{t}$, te konkretizacija skupa $\boldsymbol{\Theta}$ dopuštenih vrijednosti za parametar $\boldsymbol{t}$.

Osnovna podjela regresijskih modela jest na ***linearne*** i ***nelinearne modele***.

2. Višedimenzionalna linearna regresija

Najbolje proučeni model višedimenzionalne regresije jest onaj u kojem se pretpostavlja regresijska funkcija oblika

$$\mu_{\boldsymbol{t}}(\boldsymbol{x}) = a_1 x^{(1)} + a_2 x^{(2)} + \ldots + a_r x^{(r)} , \tag{5}$$

gdje su $a_1, \ldots, a_r$ konstantni realni brojevi (koeficijenti), pa se može reći da je $\boldsymbol{t} = \boldsymbol{a} = (a_1, \ldots, a_r) \in \mathbf{R}^r$ vektorski parametar dimenzije r.

Koristeći se pojmom skalarnog produkta vektora, možemo pisati

$$\mu_{\boldsymbol{t}}(\boldsymbol{x}) = \mu_{\boldsymbol{a}}(\boldsymbol{x}) = \boldsymbol{a}\, \boldsymbol{x}^{\mathsf{T}} . \tag{6}$$

gdje $\boldsymbol{a}\boldsymbol{x}$ označava skalarni produkt vektora $\boldsymbol{a}$ i $\boldsymbol{x}$. Vektori $\boldsymbol{a} \in \mathbf{R}^r$ i $\boldsymbol{x} \in \mathbf{R}^r$ tretiraju se kao jednoredne matrice tipa $1 \times r$, pa oznaka $\boldsymbol{x}^{\mathsf{T}}$ označuje transponiranu matricu od $\boldsymbol{x}$, a skalarno množenje vektora poistovjećuje se s matričnim množenjem.

Pretpostavimo da je $n > r$ i uvedimo još i oznake

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n) , \qquad \boldsymbol{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n) , \tag{7}$$

$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(r)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(r)} \\ \vdots & \vdots & \vdots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(r)} \end{bmatrix} . \tag{8}$$

Iz (7) se razabire da su Y (*vektor izlaznih podataka*) i $\mathcal{E}$ (*vektor greške*) n-dimenzionalni slučajni vektori s komponentama Y_i, odnosno $\mathcal{E}_i$ $(i = 1, \dots, n)$, a iz (8) se vidi da je $\boldsymbol{X}$ matrica sa n redaka i r stupaca, tj. tipa $n \times r$, koja se zove *matrica ulaznih podataka*.

Primjenom uvedenih oznaka regresijski model, s regresijskom funkcijom (5), može se zapisati u matričnom obliku

(9) $$Y = \boldsymbol{a}\boldsymbol{X}^{\mathsf{T}} + \mathcal{E} \, ,$$

gdje je $\boldsymbol{X}^{\mathsf{T}}$ transponirana matrica od $\boldsymbol{X}$, a za slučajni vektor $\mathcal{E}$ vrijedi

(10) $$\mathrm{E}[\mathcal{E}] = \boldsymbol{O} \, , \quad \boldsymbol{\Sigma}_{\mathcal{E}} = \sigma^2 \boldsymbol{I}_n \, .$$

($\boldsymbol{\Sigma}_{\mathcal{E}}$ je kovarijancna matrica s.vk. $\mathcal{E}$).

Time je izrečeno da je vektor očekivanja $\mathrm{E}[\mathcal{E}] = (\mathrm{E}[\mathcal{E}_1], \dots, \mathrm{E}[\mathcal{E}_n])$ nul-vektor dimenzije n ($\boldsymbol{O}$ je oznaka za nul-vektor), dok je kovarijancna matrica $\boldsymbol{\Sigma}_{\mathcal{E}}$ dijagonalna kvadratna matrica reda n s članom σ^2 duž glavne dijagonale ($\boldsymbol{I}_n$ je oznaka za jediničnu matricu reda n). U jednadžbi (9) vektori se tretiraju kao jednoredne matrice, tako da su Y i $\mathcal{E}$ matrice tipa $1 \times n$, dok je $\boldsymbol{a}$ matrica tipa $1 \times r$.

Time je definiran ***r*-dimenzionalni linearni regresijski model**.

Odmah primijetimo da se $(r-1)$-dimenzionalni $(r \geq 2)$ regresijski model s regresijskom funkcijom oblika

(11) $$\mu_{\boldsymbol{a}}(\boldsymbol{x}) = a_1 x^{(1)} + \cdots + a_{r-1} x^{(r-1)} + a_r$$

može tretirati kao r-dimenzionalni linearni regresijski model u kojem se uzima $x_i^{(r)} = 1$ $(i = 1, \dots, n)$, tj. r-ti stupac matrice $\boldsymbol{X}$ iz (8) sastoji se od samih jedinica.

Sada se može pristupiti rješavanju glavnog problema, tj. određivanju procjene $\widehat{\boldsymbol{t}} = \widehat{\boldsymbol{a}}$ za nepoznati parametar $\boldsymbol{a}$, koja zadovoljava uvjet (4). Niz vrijednosti regresijske funkcije

$$(\mu_{\boldsymbol{a}}(\boldsymbol{x}_1), \dots, \mu_{\boldsymbol{a}}(\boldsymbol{x}_n)) = \mu_{\boldsymbol{a}}(\boldsymbol{X}) = \boldsymbol{a}\boldsymbol{X}$$

može se, dakako, shvatiti i kao vektor iz prostora $\mathbf{R}^n$, ovisan o vektorskom parametru (vektoru koeficijenata) $\boldsymbol{a}$. Budući da je i $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbf{R}^n$, veličina

(12) $$\|\boldsymbol{y} - \mu_{\boldsymbol{a}}(\boldsymbol{X})\| = \sqrt{\sum_{i=1}^{n} [y_i - \mu_{\boldsymbol{a}}(\boldsymbol{x}_i)]^2}$$

izražava apstraktnu udaljenost (*euklidsku distancu*) između vektora $\boldsymbol{y}$ i $\mu_{\boldsymbol{a}}(\boldsymbol{X})$ u prostoru $\mathbf{R}^n$, pa se problem formuliran u (4) može shvatiti kao određivanje onoga vektora (vektorskog parametra) $\boldsymbol{a} = \widehat{\boldsymbol{a}} \in \mathbf{R}^r$ za koji su vektor izlaznih podataka $\boldsymbol{y} \in \mathbf{R}^n$ i vektor $\mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{X})$ (ovisan o ulaznim podacima) najbliži, tj. njihova euklidska distanca je najmanja.

Shvati li se svaki stupac matrice $\boldsymbol{X}$ iz (8) kao n-člani niz $(x_1^{(j)}, \dots, x_n^{(j)}) = \boldsymbol{x}^{(j)}$, tj. kao vektor iz prostora $\mathbf{R}^n$ $(\boldsymbol{x}^{(j)} \in \mathbf{R}^n, \ j = 1, \dots, r)$, može se pisati

(13) $$\mu_{\boldsymbol{a}}(\boldsymbol{X}) = \boldsymbol{a}\boldsymbol{X}^{\mathsf{T}} = a_1 \boldsymbol{x}^{(1)} + \cdots + a_r \boldsymbol{x}^{(r)} \, ,$$

što znači da se svaki vektor $\mu_{\boldsymbol{a}}(\boldsymbol{X}) \in \mathbf{R}^n$ $(\boldsymbol{a} \in \mathbf{R}^r)$ može prikazati kao linearna kombinacija vektora $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)} \in \mathbf{R}^n$. Ako su vektori $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)}$ linearno nezavisni, onda oni razapinju r-dimenzionalni podprostor L_r u prostoru $\mathbf{R}^n$, pa se problem iz (4) može formulirati i kao traženje onog vektora $\mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{X})$ u potprostoru L_r (hiperravnini u $\mathbf{R}^n$) koji je najbliži vektoru izlaznih podataka $\boldsymbol{y}$. Geometrijski zor nas upućuje na to da je to onaj vektor $\mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{X}) \in L_r$ koji se dobiva kao ortogonalna projekcija vektora $\boldsymbol{y}$ na hiperravninu L_r (v. sl. 34). Tada je vektor

$$\boldsymbol{y} - \mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{X}) = \boldsymbol{y} - \widehat{\boldsymbol{a}}\boldsymbol{X}^{\mathsf{T}} \in \mathbf{R}^n$$

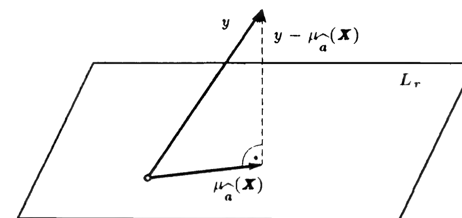
okomit na svaki vektor potprostora L_r, pa mora vrijediti

$$(\boldsymbol{y} - \widehat{\boldsymbol{a}}\boldsymbol{X}^{\mathsf{T}})\boldsymbol{x}^{(j)} = 0 \, , \qquad j = 1, \dots, r \, ,$$

što se u matričnom obliku zapisuje kao

(14) $$(\boldsymbol{y} - \widehat{\boldsymbol{a}}\boldsymbol{X}^{\mathsf{T}})\boldsymbol{X} = \boldsymbol{O} \, ,$$

gdje je $\boldsymbol{O}$ nul-vektor dimenzije r.



Slika 34. Skica odnosa vektora $\boldsymbol{y}$, $\mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{X})$ i $\boldsymbol{y} - \mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{X})$

Budući da su po pretpostavci $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(r)}$ linearno nezavisni vektori, što zapravo znači da između ulaznih varijabli $x^{(1)}, \dots, x^{(r)}$ ne postoji linearna zavisnost (ne može se nijedna od njih eliminirati tako da se izrazi kao linearna kombinacija preostalih), to je simetrična kvadratna matrica r-tog reda $\boldsymbol{B} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ regularna matrica $(\det \boldsymbol{B} \neq 0)$, što znači da postoji inverzna matrica $\boldsymbol{B}^{-1}$. Jednadžba (14) može se sada pisati

(15) $$\boldsymbol{y}\boldsymbol{X} - \widehat{\boldsymbol{a}}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} = \boldsymbol{y}\boldsymbol{X} - \widehat{\boldsymbol{a}}\boldsymbol{B} = \boldsymbol{O} \, ,$$

odnosno

(16) $$\widehat{a} = y\boldsymbol{X}\boldsymbol{B}^{-1} = y\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\,.$$

Formulom (16) riješen je problem nalaženja MNK-procjene $\widehat{\boldsymbol{a}}$ vektorskog parametra $\boldsymbol{a}$ r-dimenzionalnoga linearnog regresijskog modela (9).

Zanimljivo je da se isti rezultat dobiva (v. zad. 4) primjenom diferencijalnog računa na određivanje minimuma funkcije

$$\boldsymbol{a} \mapsto \sum_{i=1}^{n}[y_i - \mu_{\boldsymbol{a}}(\boldsymbol{x}_i)]^2 = \sum_{i=1}^{n}(y_i - a_1x_i - \ldots - a_rx_i)^2\,, \qquad \boldsymbol{a} \in \mathbf{R}^r\,,$$

pri čemu se problem svodi na rješavanje sustava jednadžbi

(17) $$\frac{\partial}{\partial a_j}\left[\sum_{i=1}^{n}(y_i - a_1x_i^{(1)} - \ldots - a_rx_i^{(r)})^2\right] = 0\,, \qquad j = 1,\ldots,r\,.$$

Tako smo, zapravo, postupili u XI.1, gdje je obrađen regresijski model koji se može shvatiti kao specijalni slučaj r-dimenzionalnoga linearnog regresijskog modela za $r = 2$ (v. zad. 2).

Primijenimo li linearni regresijski model na problem iz 1. primjera, vidimo da je $r = 4$, $n = 7$, $\boldsymbol{y} = (29{,}0\,;\ 30{,}5\,;\ 33{,}8\,;\ 37{,}5\,;\ 39{,}8\,;\ 41{,}3\,;\ 42{,}3)$ i

$$\boldsymbol{X} = \begin{bmatrix} 200 & 2123 & 0{,}600 & 1 \\ 220 & 2090 & 0{,}560 & 1 \\ 250 & 2060 & 0{,}510 & 1 \\ 280 & 2014 & 0{,}470 & 1 \\ 300 & 1993 & 0{,}450 & 1 \\ 320 & 1967 & 0{,}430 & 1 \\ 350 & 1926 & 0{,}405 & 1 \end{bmatrix}.$$

Sada se može izračunati:

$$\boldsymbol{B} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} = \begin{bmatrix} 544\,200 & 3\,865\,000 & 917 & 1\,920 \\ 3\,865\,000 & 28\,731\,000 & 6964 & 14\,170 \\ 917 & 6\,964 & 1{,}7 & 3{,}4 \\ 1\,920 & 14\,170 & 3{,}4 & 7 \end{bmatrix},$$

a zatim i

$$\boldsymbol{B}^{-1} = \begin{bmatrix} 0{,}03 & 0{,}021 & 1{,}634 & -51{,}993 \\ 0{,}021 & 0{,}017 & -0{,}186 & -39{,}467 \\ 1{,}643 & -0{,}186 & 1446 & -779{,}222 \\ -51{,}993 & -39{,}467 & -779{,}222 & 94\,550 \end{bmatrix}$$

te, prema (16), i

$$\widehat{a} = y\boldsymbol{X}\boldsymbol{B}^{-1} = (0{,}111\,;\ 0{,}058\,;\ -47{,}28\,;\ -89{,}022)\,.$$

Vidimo da su

$$\widehat{a}_1 = 0{,}111\,, \quad \widehat{a}_2 = 0{,}058\,, \quad \widehat{a}_3 = -47{,}28\,, \quad \widehat{a}_4 = -89{,}022$$

MNK-procjene regresijskih koeficijenata a_1, a_2, a_3 i a_4 u linearnome regresijskom modelu, tako da pripadna procjena regresijske funkcije glasi

$$\mu_{\widehat{\boldsymbol{a}}}(x^{(1)}, x^{(2)}, x^{(3)}) = 0{,}111\,x^{(1)} + 0{,}058\,x^{(2)} - 47{,}28\,x^{(3)} - 89{,}022\,.$$

3. Gauss-Markovljev teorem

Već su pri razmatranju jednodimenzionalnoga regresijskog modela istaknuta dobra svojstva metode najmanjih kvadrata za procjenu nepoznatih parametara, što naravno vrijedi i za višedimenzionalnu linearnu regresiju. Posebno je važno svojstvo koje se obično izriče kao *Gauss-Markovljev teorem.*

Primijetimo najprije da iz definicije r-dimenzionalnoga linearnog regresijskog modela slijedi da se vektor $\widehat{\boldsymbol{a}}$ iz (16) može shvatiti kao vrijednost vektorskog procjenitelja $\widehat{\boldsymbol{A}} = (\widehat{A}_1, \ldots, \widehat{A}_r)$ za nepoznati vektorski parametar $\boldsymbol{a} = (a_1, \ldots, a_r)$, pa se može pisati

(18) $$\widehat{\boldsymbol{A}} = \boldsymbol{Y}\boldsymbol{X}\boldsymbol{B}^{-1} = \boldsymbol{Y}\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\,,$$

iz čega se razabire da je procjenitelj $\widehat{\boldsymbol{A}}$ izražen kao linearna funkcija izlaznoga slučajnog vektora $\boldsymbol{Y}$. To, nadalje, znači da se može pisati

(19) $$\widehat{A}_j = \sum_{i=1}^{n}\beta_{ij}Y_i\,, \qquad j = 1,\ldots,r\,,$$

gdje je β_{ij} element u i-tom retku i j-tom stupcu matrice $\boldsymbol{X}\boldsymbol{B}^{-1}$, i to je neslučajna veličina. Zato se može reći da je slučajna varijabla $\widehat{A}_j$ izražena kao linearna kombinacija slučajnih varijabli $Y_1, \ldots, Y_n$ s koeficijentima $\beta_{ij}(i = 1, \ldots, n)$. To za $\widehat{A}_j$ opravdava naziv *linearni procjenitelj*, jer se iz (19) vidi da je statistika $\widehat{A}_j$ linearna funkcija slučajnog uzorka. Također se za slučajni vektor $\widehat{\boldsymbol{A}}$ kaže da je *linearni procjenitelj za vektorski parametar* $\boldsymbol{a}$.

Sada se, naravno, postavlja zadatak da se ustanovi veza između vektora očekivanja $\mathrm{E}[\widehat{\boldsymbol{A}}] = (\mathrm{E}[A_1], \ldots, \mathrm{E}[A_r])$ i $\mathrm{E}[\boldsymbol{Y}] = (\mathrm{E}[Y_1], \ldots, [Y_n])$, te odgovarajućih kovarijancnih matrica $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{A}}}$ i $\boldsymbol{\Sigma}_{\boldsymbol{Y}}$ (v. V.5). Uzme li se u obzir (7), (8), (9) i (18), može se pisati

$$\mathrm{E}[\widehat{\boldsymbol{A}}] = \mathrm{E}[(\boldsymbol{a}\boldsymbol{X}^{\mathsf{T}} + \boldsymbol{\mathcal{E}})\boldsymbol{X}\boldsymbol{B}^{-1}] = \boldsymbol{a}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{B}^{-1} + \mathrm{E}[\boldsymbol{\mathcal{E}}]\boldsymbol{X}\boldsymbol{B}^{-1}\,,$$

što zbog $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} = \boldsymbol{B}$ i (10) postaje

(20) $$\mathrm{E}[\widehat{\boldsymbol{A}}] = \boldsymbol{a}\,,$$

pa se može reći da je slučajni vektor $\widehat{\boldsymbol{A}} = (\widehat{A}_1, \ldots, \widehat{A}_r)$ nepristrani procjenitelj za vektorski parametar $\boldsymbol{a} = (a_1, \ldots, a_r)$.

Vidimo da je slučajni vektor $\widehat{\boldsymbol{A}}$ linearan i nepristran procjenitelj za vektorski parametar $\boldsymbol{a}$, pa se kaže da je on LN-*procjenitelj*.

Polazeći od (9) i (10), te formula navedenih u zad. 22. iz V. poglavlja odmah se vidi da je

(21) $$\boldsymbol{\Sigma}_{\boldsymbol{Y}} = \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}}} = \sigma^2 \boldsymbol{I}_n \,.$$

Primjenjujući, pak, formule iz zad. 23. u V. poglavlju, na slučajni vektor $\widehat{\boldsymbol{A}}$ iz (18), dobiva se

(22) $$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{A}}} = (\boldsymbol{X}\,\boldsymbol{B}^{-1})^{\mathsf{T}}\,\boldsymbol{\Sigma}_{\boldsymbol{Y}}\,(\boldsymbol{X}\,\boldsymbol{B}^{-1}) = \sigma^2 \boldsymbol{B}^{-1} \,,$$

jer je $(\boldsymbol{X}\,\boldsymbol{B}^{-1})^{\mathsf{T}} = (\boldsymbol{B}^{-1})^{\mathsf{T}}\,\boldsymbol{X}^{\mathsf{T}} = (\boldsymbol{B}^{\mathsf{T}})^{-1}\,\boldsymbol{X}^{\mathsf{T}} = \boldsymbol{B}^{-1}\,\boldsymbol{X}^{\mathsf{T}}$ $(\boldsymbol{B} = \boldsymbol{B}^{\mathsf{T}}$, jer je $\boldsymbol{B}$ simetrična matrica).

Iz (22) se vidi da je varijanca procjenitelja $\widehat{A}_j$ nepoznatog parametra a_j

(23) $$\mathrm{D}[\widehat{A}_j] = \sigma^2\, b_{jj} \,, \qquad j = 1, \ldots, r \,,$$

gdje je b_{jj} dijagonalni element matrice $\boldsymbol{B}^{-1}$.

Odmah se može postaviti pitanje da li je procjenitelj $\widehat{A}_j$ najbolji, u smislu da ima najmanju varijancu, u klasi svih LN-procjenitelja za nepoznati parametar a_j. Odgovor na to pitanje, i još više, daje nam Gauss-Markovljev teorem (GM-*teorem*).

Pretpostavke:

1. $\boldsymbol{Y} = \boldsymbol{a}\boldsymbol{X}^{\mathsf{T}} + \boldsymbol{\mathcal{E}}$ je r-dimenzionalni linearni regresijski model za koji je $\boldsymbol{B} = \boldsymbol{X}^{\mathsf{T}}\,\boldsymbol{X}$ regularna matrica.
2. $\widehat{\boldsymbol{A}} = (A_1, \ldots, A_r)$ je MNK-procjenitelj za nepoznati vektorski parametar $\boldsymbol{a} = (a_1, \ldots, a_r)$.
3. $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_r) \in \mathbf{R}^r$ je proizvoljni r-dimenzionalni vektor.
4. $\mu_{\boldsymbol{a}}(\boldsymbol{\xi}) = \boldsymbol{a}\,\boldsymbol{\xi}^{\mathsf{T}} = a_1\xi_1 + \ldots + a_r\xi_r$ je vrijednost regresijske funkcije u točki $\boldsymbol{\xi} \in \mathbf{R}^r$.

Tvrdnja:

Slučajna varijabla (statistika)

(24) $$\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{\xi}) = \widehat{\boldsymbol{A}}\,\boldsymbol{\xi}^{\mathsf{T}} = \widehat{A}_1\xi_1 + \ldots + \widehat{A}_r\xi_r$$

je ***najbolji linearni nepristrani procjenitelj*** (**NLN-*procjenitelj***) za veličinu $\mu_{\boldsymbol{a}}(\boldsymbol{\xi})$ u smislu da je

(25) $$\mathrm{V}[\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{\xi})] = \mathrm{V}[\widehat{\boldsymbol{A}}\,\boldsymbol{\xi}^{\mathsf{T}}] \leq \mathrm{V}[\widehat{\boldsymbol{T}}\,\boldsymbol{\xi}^{\mathsf{T}}]$$

za svako $\boldsymbol{\xi} \in \mathbf{R}^r$ i svaki LN-procjenitelj $\widehat{\boldsymbol{T}} = (\widehat{T}_1, \ldots, \widehat{T}_r)$ za vektorski parametar $t = \boldsymbol{a}$.

Uzme li se, na primjer, $\boldsymbol{\xi} = (1, 0, \ldots, 0)$, dobiva se $\mu_{\boldsymbol{a}}(\boldsymbol{\xi}) = a_1$ i GM-teorem tvrdi da je $\widehat{A}_1$ nepristrani linearni procjenitelj za parametar a_1, s najmanjom varijancom, tj. NLN-procjenitelj.

Slično se može uočiti da je $\widehat{A}_j$ $(j = 1, \ldots, r)$ NLN-procjenitelj za parametar a_j.

Očigledno je da je $\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{\xi})$ iz (24) linearan procjenitelj, a odmah ćemo dokazati da je i nepristran. To, naime, slijedi neposredno iz (20) i formule u zad. 23a iz V. poglavlja, jer je

(26) $$\mathrm{E}[\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{\xi})] = \mathrm{E}[\widehat{\boldsymbol{A}}\,\boldsymbol{\xi}^{\mathsf{T}}] = \mathrm{E}[\widehat{\boldsymbol{A}}]\,\boldsymbol{\xi}^{\mathsf{T}} = \boldsymbol{a}\,\boldsymbol{\xi}^{\mathsf{T}} = \mu_{\boldsymbol{a}}(\boldsymbol{\xi}) \,.$$

Ostalo je još da se dokaže valjanost relacije (25). Linearnost procjenitelja $\widehat{\boldsymbol{T}}$ izražava se jednadžbom

(27) $$\widehat{\boldsymbol{T}} = \boldsymbol{Y}\,\boldsymbol{C} \,,$$

gdje je $\boldsymbol{C}$ proizvoljna realna matrica tipa $n \times r$. Uvjet nepristranosti vektorskog procjenitelja $\widehat{\boldsymbol{T}}$, tj. jednadžba $\mathrm{E}[\widehat{\boldsymbol{T}}] = \boldsymbol{a}$, može se, imajući na umu (9) i (10), zapisati i kao

$$\mathrm{E}[\boldsymbol{Y}\,\boldsymbol{C}] = \mathrm{E}[\boldsymbol{Y}]\,\boldsymbol{C} = \mathrm{E}[\boldsymbol{a}\,\boldsymbol{X}^{\mathsf{T}} + \boldsymbol{\mathcal{E}}]\,\boldsymbol{C} = \boldsymbol{a}\,\boldsymbol{X}^{\mathsf{T}}\,\boldsymbol{C} = \boldsymbol{a} \,,$$

odnosno kao

(28) $$\boldsymbol{X}^{\mathsf{T}}\,\boldsymbol{C} = \boldsymbol{I}_r \,,$$

gdje je $\boldsymbol{I}_r$ jedinična matrica r-tog reda.

Cilj nam je dokazati nejednakost

$$\mathrm{V}[\widehat{\boldsymbol{T}}\,\boldsymbol{\xi}^{\mathsf{T}}] - \mathrm{V}[\widehat{\boldsymbol{A}}\,\boldsymbol{\xi}^{\mathsf{T}}] \geq 0 \,.$$

To će se postići na taj način što će se dokazati da je $\mathrm{V}[\widehat{\boldsymbol{T}}\,\boldsymbol{\xi}^{\mathsf{T}}] - \mathrm{V}[\widehat{\boldsymbol{A}}\,\boldsymbol{\xi}^{\mathsf{T}}]$ varijanca slučajne varijabla $\boldsymbol{Z}\,\boldsymbol{\xi}^{\mathsf{T}}$, gdje je $\boldsymbol{Z} = \widehat{\boldsymbol{T}} - \widehat{\boldsymbol{A}} = \boldsymbol{Y}\,(\boldsymbol{C} - \boldsymbol{X}\,\boldsymbol{B}^{-1})$. Stavimo

(29) $$\boldsymbol{Q} = \boldsymbol{C} - \boldsymbol{X}\,\boldsymbol{B}^{-1} \Rightarrow \boldsymbol{C} = \boldsymbol{Q} + \boldsymbol{X}\,\boldsymbol{B}^{-1} \,,$$

pa je

$$\boldsymbol{Z} = \boldsymbol{Y}\,\boldsymbol{Q} \,.$$

Primjenom formule iz zad. 24. u V. poglavlju dobiva se

(30) $$\left\{ \begin{aligned} \mathrm{V}[\widehat{\boldsymbol{A}}\,\boldsymbol{\xi}^{\mathsf{T}}] &= \boldsymbol{\xi}\,\boldsymbol{\Sigma}_{\widehat{\boldsymbol{A}}}\,\boldsymbol{\xi}^{\mathsf{T}} \,, \\ \mathrm{V}[\widehat{\boldsymbol{T}}\,\boldsymbol{\xi}^{\mathsf{T}}] &= \boldsymbol{\xi}\,\boldsymbol{\Sigma}_{\widehat{\boldsymbol{T}}}\,\boldsymbol{\xi}^{\mathsf{T}} \,, \\ \mathrm{V}[\boldsymbol{Z}\,\boldsymbol{\xi}^{\mathsf{T}}] &= \boldsymbol{\xi}\,\boldsymbol{\Sigma}_{\boldsymbol{Z}}\,\boldsymbol{\xi}^{\mathsf{T}} \,. \end{aligned} \right.$$

Kovarijancna matrica $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{A}}}$ izražena je u (22), pa ostaje da se nađu izrazi za $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{T}}}$ i $\boldsymbol{\Sigma}_{\boldsymbol{Z}}$. Iz (27) i već spomenute formule u 23.b zadatku u V. poglavlju, proizlazi

(31) $$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{T}}} = \boldsymbol{C}^{\mathsf{T}}\,\boldsymbol{\Sigma}_{\boldsymbol{Y}}\,\boldsymbol{C} = (\text{ zbog } (21)) = \sigma^2 \boldsymbol{C}^{\mathsf{T}}\,\boldsymbol{C} \,.$$

Slično se izvodi

(32) $$\boldsymbol{\Sigma}_Z = \boldsymbol{Q}^\top \boldsymbol{\Sigma}_Y \boldsymbol{Q} = \sigma^2 \boldsymbol{Q}^\top \boldsymbol{Q} = \sigma^2(\boldsymbol{C}^\top \boldsymbol{C} - \boldsymbol{B}^{-1}).$$

Transponiranjem jednadžbe (28) dobiva se, naime, da je $\boldsymbol{C}^\top \boldsymbol{X} = \boldsymbol{I}_r$, a iz (29) proizlazi da je

$$\boldsymbol{Q}^\top \boldsymbol{Q} = (\boldsymbol{C}^\top - \boldsymbol{B}^{-1}\boldsymbol{X}^\top)(\boldsymbol{C} - \boldsymbol{X}\boldsymbol{B}^{-1}) =$$
$$= \boldsymbol{C}^\top \boldsymbol{C} - \boldsymbol{B}^{-1}(\boldsymbol{X}^\top \boldsymbol{C}) - (\boldsymbol{C}^\top \boldsymbol{X})\boldsymbol{B}^{-1} + \boldsymbol{B}^{-1}(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{B}^{-1} = \boldsymbol{C}^\top \boldsymbol{C} - \boldsymbol{B}^{-1}.$$

Na temelju (22), (30), (31) i (32) može se pisati

$$\mathrm{V}[\widehat{T}\xi^\top] - \mathrm{V}[\widehat{A}\xi^\top] = \xi\left(\boldsymbol{\Sigma}_{\widehat{\boldsymbol{T}}} - \boldsymbol{\Sigma}_{\widehat{\boldsymbol{A}}}\right)\xi^\top =$$
$$= \xi[\sigma^2(\boldsymbol{C}^\top \boldsymbol{C} - \boldsymbol{B}^{-1})]\xi^\top = \xi\,\boldsymbol{\Sigma}_Z\,\xi^\top = \mathrm{V}[Z\xi^\top] \geq 0,$$

i time je u potpunosti dokazan GM-teorem.

Iz prve jednadžbe u (30), te (22) i (25) odmah se dobiva formula

(33) $$\mathrm{V}[\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{\xi})] = \sigma^2 \boldsymbol{\xi}\boldsymbol{B}^{-1}\boldsymbol{\xi}^\top,$$

koja pokazuje ovisnost varijance NLN-procjenitelja $\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{\xi})$ za nepoznatu vrijednost $\mu_a(\xi)$ regresijske funkcije u točki $\boldsymbol{\xi} \in \mathbf{R}^r$ o nepoznatom parametru σ^2. Stoga se prirodno nameće zadaća da se pronađe dobar procjenitelj za parametar σ^2. Može se dokazati (v. XII.6) da je statistika

(34) $$S^2 = \frac{1}{n-r}\sum_{i=1}^{n}[Y_i - \mu_{\widehat{\boldsymbol{A}}}(x_i)]^2$$

nepristrani procjenitelj za nepoznati parametar σ^2, tj. da vrijedi

(35) $$\mathrm{E}[S^2] = \sigma^2.$$

4. Tablica analize varijance

Kao što smo u XI.3. opisali analizu rasipanja izlaznih podataka u jednodimenzionalnome linearnom regresijskom modelu, može se isto učiniti i u r-dimenzionalnome linearnom regresijskom modelu. Uveđu li se oznake

(36) $$\widehat{y}_i = \widehat{\mu}_i = \mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{x}_i), \qquad \widehat{\varepsilon}_i = y_i - \widehat{y}_i, \qquad i = 1, \ldots, n,$$

reći ćemo da je $\widehat{y}_i$ procjena vrijednosti izlazne slučajne varijable u točki $\boldsymbol{x}_i \in \mathbf{R}^r$, dok je $\widehat{\varepsilon}_i$ reziduum, sa značenjem razlike između izmjerene i procijenjene vrijednosti izlazne varijable u toj točki. Zato se veličina

$$\sum_{i=1}^{n}\widehat{\varepsilon}_i^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = n\widehat{\sigma}^2$$

zove *zbroj kvadrata grešaka*, a veličina

$$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = n\widehat{\sigma}_0^2$$

zove se *regresijski zbroj kvadrata*, dok se veličina

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = ns_y^2$$

zove *ukupni zbroj kvadrata* odstupanja od aritmetičke sredine izlaznih podataka $(\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i)$.

Opet se dokazuje (v. zad.6) da je

(37) $$s_y^2 = \widehat{\sigma}_0^2 + \widehat{\sigma}^2,$$

što omogućuje da se definira *koeficijent determinacije*

(38) $$R^2 = \frac{\widehat{\sigma}_0^2}{s_y^2} = 1 - \frac{\widehat{\sigma}^2}{s_y^2},$$

sa značenjem opisanim u XI.3.

Ovako definiran koeficijent determinacije može dati i krivu sliku o stvarnom odnosu rasipanja uzrokovanih regresijskom ovisnošću i slučajnom greškom, posebno kada broj n podataka nije mnogo veći od broja r kontroliranih varijabli. Kada bismo, na primjer, imali $n = 2$ i $r = 2$, tj. dvije različite točke u ravnini, tada je jasno da će kroz te točke prolaziti procijenjeni regresijski pravac, tako da će biti $\widehat{\sigma}^2 = 0$, pa stoga $R^2 = 1$, što bismo interpretirali tako da je ukupno rasipanje prouzročeno regresijskom funkcijom. No, očigledno je presmiono, na temelju samo dva podatka, uzeti dobiveni pravac kao stvarnu krivulju regresije. Zato se, pri relativno malom broju podataka n, u odnosu na dimenziju r regresijskog modela, definira tzv. *korigirani koeficijent determinacije*

(39) $$\overline{R}^2 = 1 - \frac{\frac{1}{n-r}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{(n-1)\widehat{\sigma}^2}{(n-r)s_y^2}.$$

Da bismo objasnili definicijsku formulu (39) primijetimo da se ona razlikuje od definicijske formule (38) po tome, što su umjesto vrijednosti $\widehat{\sigma}^2$ i s_y^2 pristranih procjenitelja za nepoznate parametre σ^2 i σ_y^2 (σ_y^2 je teorijska vrijednost varijance izlazne varijable) uvrštene vrijednosti $\frac{1}{n-r}\widehat{\sigma}^2 = s^2$ i $\frac{1}{n-1}s_y^2$ odgovarajućih

nepristranih procjenitelja. Govori se da postoji $n-r$ stupnjeva slobode za formiranje rezidualnog rasipanja $\widehat{\sigma}^2$, jer je potrebno r različitih točaka (podataka) za određivanje r nepoznatih parametara (koeficijenata) regresijske funkcije, pa ostaje $n-r$ slobodnih podataka za oblikovanje rezidualnog rasipanja. Na sličan način se objašnjava da za formiranje ukupnog rasipanja (s_y^2) postoji $n-1$ stupnjeva slobode, jer od ukupno n podataka ostaje slobodno njih $n-1$ za formiranje rasipanja oko aritmetičke sredine $\overline{y}$.

Za podatke iz 1. primjera dobili bismo $s_y^2 = 23{,}98$ i $\widehat{\sigma}^2 = 0{,}31$, pa koeficijent determinacije, prema (38), iznosi $R^2 = 0{,}987$, dok korigirani koeficijent determinacije, prema (39), iznosi $\overline{R}_2 = 0{,}974$.

Dobiveni rezultat interpretira se tako da se kaže da je 97,4 % rasipanja izlaznih podataka uzrokovano regresijskom funkcijom, dok se samo 2,6 % preostalog (rezidualnog) rasipanja ne može objasniti usvojenim regresijskim modelom. To bi upućivalo na dobru prilagodbu modela danim empirijskim podacima.

Pregledan i jasan uvid u rasipanje izlaznih podataka, slično kao i u jednodimenzionalnom modelu (v. tabl. 2. u XI.4), omogućuje tablica analize varijance.

Tablica 2.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Srednje kvadratno odstupanje	Koeficijent determinacije (korigirani)
model	$r-1$	$\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2$	$\widehat{\sigma}_0^2$	
slučajna greška	$n-r$	$\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2$	$\widehat{\sigma}^2$	
ukupno rasipanje	$n-1$	$\sum_{i=1}^{n}(y_i-\overline{y})^2$	s_y^2	R^2 ($\overline{R}^2$)

Za konkretne podatke iz 1. primjera rezultati su prikazani u tabl. 3.

Tablica 3.

Model	3	165,71	23,67	
Slučajna greška	3	2,16	0,31	
Ukupno rasipanje	6	167,87	23,98	0,987 (0,974)

5. Intervali povjerenja za regresijske koeficijente

Primijetimo najprije da su svi dosadašnji rezultati u vezi s r-dimenzionalnom linearnom regresijom izvedeni na temelju modela izraženog formulom (9) i uz tzv. *standardne pretpostavke*:

1. Matrica ulaznih podataka $\boldsymbol{X}$ ima rang r ($r < n$), što znači da ne postoji linearna zavisnost između ulaznih varijabli.
2. Vektor očekivanja slučajnog vektora $\boldsymbol{\mathcal{E}}$ je nul-vektor, što znači da je očekivana vrijednost izlazne varijable jednaka vrijednosti regresijske funkcije.
3. Disperzijska matrica slučajnog vektora $\boldsymbol{\mathcal{E}}$ dijagonalna je matrica s članom σ^2 duž glavne dijagonale, što znači da su greške nekorelirane slučajne varijable sa zajedničkom disperzijom σ^2.

Da bi se dobili i određeni rezultati, pomoću kojih se ocjenjuju greške pri procjeni nepoznatih parametara r-dimenzionalnoga linearnog regresijskog modela, potrebno je usvojiti i dodatnu pretpostavku:

4. Slučajni vektor $\boldsymbol{\mathcal{E}}$ podvrgava se n-dimenzionalnoj normalnoj radiobi (v. V.5), kojoj je n-dimenzionalni nul-vektor $\boldsymbol{O}$ vektor očekivanja i $\sigma^2\boldsymbol{I}_n$ kovarijancna matrica.

Iz te dodatne pretpostavke i formule (9) odmah slijedi da slučajnom vektoru $\boldsymbol{Y}$ pripada n-dimenzionalna normalna razdioba s vektorom $\mu_{\boldsymbol{a}}(\boldsymbol{X}) = \boldsymbol{a}\boldsymbol{X}^{\mathrm{T}}$ kao vektorom očekivanja i matricom $\sigma^2\boldsymbol{I}_n$ kao kovarijancnom matricom.

Formula (18) pokazuje, pak, da je vektorski procjenitelj $\widehat{\boldsymbol{A}}$, za nepoznati parametar $\boldsymbol{a}$, promatranog r-dimenzionalnoga linearnog regresijskog modela, linearno ovisan o slučajnom vektoru $\boldsymbol{Y}$, što zajedno s dodatnom 4. pretpostavkom implicira da slučajnom vektoru $\widehat{\boldsymbol{A}}$ pripada r-dimenzionalna normalna razdioba s vektorom očekivanja $\mathrm{E}[\widehat{\boldsymbol{A}}] = \boldsymbol{a}$ (v. (20)) i kovarijancnom matricom $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{A}}} = \sigma^2\boldsymbol{B}^{-1}$ (v. (22)). To znači da komponenti $\widehat{A}_j (j = 1,\ldots,r)$ pripada normalna razdioba $\mathrm{N}(a_j, \sigma_j^2)$, gdje je $\sigma_j^2 = \sigma^2 b_{jj}$ (v. (23)).

Dobiveni rezultat pokazuje nam da vrijednost $\widehat{a}_j$, MNK-procjenitelja $\widehat{A}_j$ za nepoznati regresijski koeficijent a_j, možemo shvatiti kao vrijednost slučajne varijable koja se rasipa po normalnoj razdiobi oko a_j uz varijancu σ_j^2.

Cilj nam je odrediti interval povjerenja zadane pouzdanosti γ za nepoznati parametar a_j, ali se pritom pojavljuje teškoća zbog činjenice da σ_j^2 ovisi o nepoznatom parametru σ^2. Ako je n (broj podataka) dovoljno velik, onda se σ^2 može zamijeniti vrijednošću s^2 pripadnoga nepristranog procjenitelja S^2 iz (34). Stavimo

$$\widehat{\sigma}_j^2 = s^2 b_{jj}\,, \qquad j = 1,\ldots,r\,, \tag{40}$$

pa ćemo uzeti da približno vrijedi

$$\widehat{A}_j \sim \mathrm{N}(a_j, \widehat{\sigma}_j^2)\,,$$

iz čega proizlazi da će granice intervala povjerenja pouzdanosti γ, za nepoznati regresijski koeficijent a_j, biti izražene formulama

$$g_{1,2}^{(j)} = \widehat{a}_j \mp z_\gamma \widehat{\sigma}_j \,, \qquad j = 1,\ldots,r \,, \tag{41}$$

gdje je z_γ objašnjeno u VII.1. (tabl. 1).

Ako se ne oslanjamo na pretpostavku o velikom broju n podataka, onda ćemo poći od činjenice (v. XII.6) da

$$T_j = \frac{\widehat{A}_j - a_j}{S\, b_{jj}} \sim \mathrm{t}(n-r) \,, \qquad j = 1,\ldots,r \,, \tag{42}$$

iz čega proizlazi da su granice intervala povjerenja pouzdanosti γ, za regresijski koeficijent a_j, izražene formulama

$$g_{1,2}^{(j)} = \widehat{a}_j \mp \tau_\gamma \widehat{\sigma}_j \,, \qquad j = 1,\ldots,r \,, \tag{43}$$

gdje je τ_γ veličina koja se odnosi na Studentovu razdiobu $\mathrm{t}(n-r)$ sa $n-r$ stupnjeva slobode (v. tabl. 4. u VII.2).

Da bismo primijenili izvedene rezultate na situaciju iz 1. primjera, primijetimo najprije da iz matrice $\boldsymbol{B}^{-1}$, izračunane u XII.2, odčitavamo

$$b_{11} = 0{,}03 \,, \quad b_{22} = 0{,}017 \,, \quad b_{33} = 1446 \,, \quad b_{44} = 94\,550 \,.$$

Izračunaju li se vrijednosti $\widehat{y}_i (i = 1,\ldots,7)$, prema formuli (36), odmah se može izračunati i $s^2 = \frac{1}{3}\sum_{i=1}^{7}(y_i - \widehat{y}_i)^2 = \frac{1}{3}\,2{,}16 = 0{,}72$ $(s = 0{,}85)$.

Sada možemo, uz MNK-procjenu $\widehat{a}_j$ regresijskog koeficijenta a_j, upisati i pripadnu procjenu $\widehat{\sigma}_j$ standardne devijacije σ_j.

Tablica 4.

j	1	2	3	4
$\widehat{a}_j$	0,111	0,058	−47,28	−89,02
$\widehat{\sigma}_j$	0,146	0,109	32,2	260,7

Iz tabl. 4. vidljivo je da su uz procjene svih regresijskih koeficijenata vezane, relativno vrlo velike, standardne devijacije, što svakako upozorava na oprez pri korištenju dobivenoga regresijskog modela. Greške procjena još se bolje uočavaju pomoću odgovarajućih intervala povjerenja pouzdanosti $\gamma = 0{,}95$ $(\tau_\gamma = 3{,}18)$, izračunanih na temelju formula (43).

Očigledno je da smo dobli vrlo široke intervale povjerenja, zbog toga što je broj mjerenja $n = 7$ jako malen i stoga se mora računati s vrlo velikom mogućom greškom pri procjeni nepoznatih regresijskih koeficijenata.

Kao što se u jednodimenzionalnome regresijskom modelu prirodno pojavio problem interpolacije i ekstrapolacije regresijske funkcije, analogni se problem može

Tablica 5.

j	1	2	3	4
$g_1^{(j)}$	−0,352	−0,289	−150	−918
$g_2^{(j)}$	0,575	0,589	55,1	740

postaviti i u modelu višestruke regresije. Gauss-Markovljev teorem jamči nam da je vrijednost

$$\mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{x}) = \widehat{a}_1 x^{(1)} + \ldots + \widehat{a}_r x^{(r)} \,, \qquad \boldsymbol{x} \in \mathbf{R}^r \,, \tag{44}$$

najbolja procjena za nepoznatu vrijednost $\mu_{\boldsymbol{a}}(\boldsymbol{x})$ regresijske funkcije u r-dimenzionalnome linearnom regresijskom modelu. Formula (33) omogućuje uvid u varijancu pri spomenutoj procjeni, gdje se opet pojavljuje nepoznati parametar σ^2, koji se za velike n može zamijeniti vrijednošću s^2 pripadnoga nepristranog procjenitelja S^2 iz (34). Označimo tu varijancu sa $\widehat{\sigma}_{\boldsymbol{x}}^2$, pa se može pisati

$$\widehat{\sigma}_{\boldsymbol{x}}^2 \approx s^2\, \boldsymbol{x}\, \boldsymbol{B}^{-1}\, \boldsymbol{x}^{\mathsf{T}} \,. \tag{45}$$

Ako bismo u 1. primjeru uzeli $x^{(1)} = 230$, $x^{(2)} = 2050$, $x^{(3)} = 0{,}55$ i $x^{(4)} = 1$, primjenom (44) i već izračunanih procjena $\widehat{a}_1, \widehat{a}_2, \widehat{a}_3$ i $\widehat{a}_4$ regresijskih koeficijenata, dobili bismo procjenu

$$\mu_{\widehat{\boldsymbol{a}}}(\boldsymbol{x}) = 0{,}111 \cdot 230 + 0{,}058 \cdot 2050 - 47{,}28 \cdot 0{,}55 - 89{,}02 = 29{,}4$$

za odgovarajuću vrijednost regresijske funkcije. To je, dakako, i najbolja prognoza za vrijednost izlazne slučajne varijable, što praktički znači da možemo očekivati tlačnu čvrstoću od 29,4 MPa na betonskim kockama izrađenim po recepturi:

230 kg cementa, 2 050 kg agregata, 0,55 vodocementni faktor.

U ovom primjeru ne smijemo primijeniti formulu (45) za procjenu pripadne varijance, jer imamo premali broj podataka $(n = 7)$.

U slučaju dovoljno velikog broja n podataka može se uzeti da slučajnoj varijabli $\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{x})$, približno pripada normalna razdioba $\mathrm{N}(\mu_{\boldsymbol{a}}(\boldsymbol{x}), \widehat{\sigma}_{\boldsymbol{x}}^2)$, što omogućuje i određivanje intervala povjerenja, zadane pouzdanosti γ, za nepoznatu vrijednost regresijske funkcije $\mu_{\boldsymbol{a}}(\boldsymbol{x})$.

Relacija (42) omogućuje i konstruiranje testa za testiranje hipoteze $\mathrm{H}_0 : a_j = a_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : a_j \neq a_0$ (ili $a_j < a_0$, ili $a_j > a_0$), gdje je a_0 zadani realan broj. Slučajna varijabla T_j iz (42) uzet će se kao test-statistika, jer je očigledno da vrijednost

$$t_j = \frac{\widehat{a}_j - a_0}{s\sqrt{b_{jj}}} = \frac{a_j - a_0}{\widehat{\sigma}_j} \tag{46}$$

pokazuje odstupanje procjene $\widehat{a}_j$, dobivene na temelju danih podataka, od pretpostavljene vrijednost a_0 regresijskog koeficijenta a_j.

Kritično područje, uz alternativnu hipotezu $H_1 : a_j \neq a_0$, određeno je relacijom

$$|t_j| \geq G_{n-r}^{-1}\left(1-\frac{\alpha}{2}\right),$$

gdje je G_{n-r}^{-1} inverzna funkcija za f.r.v. hikvadrat-razdiobe sa $n-r$ stupnjeva slobode.

Uzme li se kao alternativna hipoteza $H_1 : a_j < a_0$, kritično područje određeno je nejednakošću

$$t_j \geq G_{n-r}^{-1}(1-\alpha),$$

dok je uz alternativnu hipotezu $H_1 : a_j > a_0$, kritično područje određeno nejednakošću

$$t_j \leq G_{n-r}^{-1}(\alpha).$$

Pogledamo li tablice 4. i 5. nameće nam se ideja da testiramo hipotezu $H_0 : a_1 = 0$. prema alternativnoj hipotezi $H_1 : a_1 \neq 0$, uz razinu značajnosti od, recimo $\alpha = 0{,}10$. Prema (46) dobivamo $t_1 = \frac{0{,}111-0}{0{,}146} \approx 0{,}760$, dok je $G_3^{-1}(0,95) = 2{,}353$, pa se vidi da hipotezu H_0 treba prihvatiti.

To praktički znači da nam dani podaci sugeriraju zaključak da tlačna čvrstoća betonskih kocki ne ovisi o količini cementa u primijenjenoj recepturi za proizvodnju betona!? Svaki će tehnolog betona primijetiti da tu nešto nije u redu.

No, to i nije kraj neobičnim rezultatima. Provedimo postupak testiranja hipoteze $H_0 : a_j = 0$, prema alternativnoj hipotezi $H_1 : a_j \neq 0$, za sve regresijske koeficijente $(j = 1, 2, 3, 4)$. Rezultati su prikazani u tabl. 6.

Tablica 6.

j	1	2	3	4
t_j	0,760	0,532	−1,47	−0,342

Vidimo da je u svakom slučaju $|t_j| \leq G_3^{-1}(0{,}95) = 2{,}353$ $(j = 1, 2, 3, 4)$, što znači da svaku od navedenih nul-hipoteza treba prihvatiti, jer vrijednost odgovarajuće test-statistike uvijek pada izvan kritičnog područja.

Došli smo do praktički apsurdnog zaključka da tlačna čvrstoća betonskih kocki ne ovisi ni o količini cementa, ni o količini agregata, ni o vodocementnom faktoru.

Objašnjenje se nalazi u činjenici da se nalazimo u tzv. ***nestabilnoj situaciji***, gdje je $\det \boldsymbol{B} = \det(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})$ blizu nule. To, pak, znači da gotovo postoji linearna veza među ulaznim varijablama. I zaista pogledamo li tabl. 1. vidimo da vrijedi

$$\begin{aligned} x_i^{(1)} + x_i^{(2)} &\approx 2\,300 \\ x_i^{(1)} + 100\,x_i^{(3)} &\approx 760 \end{aligned} \qquad i = 1, 2, 3, 4, 5, 6, 7.$$

No, u uvjetima postojanja linearne veze među ulaznim varijablama je $\det \boldsymbol{B} = 0$, matrica $\boldsymbol{B}$ nije regularna i ne postoji $\boldsymbol{B}^{-1}$, pa ne postoje ni MNK-procjene regresijskih koeficijenata. U nestabilnoj situaciji oni doduše postoje, ali ne odražavaju stvarni utjecaj pojedine ulazne varijable na vrijednost izlazne varijable,

iako se procijenjena regresijska funkcija dobro prilagođuje danim podacima, što se vidi preko odgovarajućeg koeficijenta determinacije (v. XII.4).

6. Fundamentalni teorem

Važan teorem, na kojem se temelje mnogi teorijski rezultati (formula (42) i dr.) i praktični postupci višestruke linearne regresije, glasi:

Ako su ispunjene standardne pretpostavke 1 – 3. i dodatna pretpostavka 4. (v. XII.5) u r-dimenzionalnome linearnom regresijskom modelu, onda slučajnoj varijabli

$$\frac{n-r}{\sigma^2}\,S^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}[Y_i - \mu_{\hat{A}}(x_i)]^2$$

pripada hikvadrat-razdioba sa $n-r$ stupnjeva slobode.

Iz ovoga odmah proizlazi da je

$$\text{(47)} \qquad \mathrm{E}[S^2] = \sigma^2, \qquad \mathrm{V}[S^2] = \frac{2}{n-r}\,\sigma^4,$$

što pokazuje da je statistika S^2 nepristran i konzistentan procjenitelj za nepoznati parametar σ^2. To opravdava postupak da se za velike n $(n \gg r)$ uzima da je

$$\text{(48)} \qquad \sigma^2 \approx s^2 = \frac{1}{n-r}\sum_{i=1}^{n}[y_i - \mu_{\hat{\boldsymbol{a}}}(x_i)]^2.$$

Primijetimo da se prva formula u (47) poklapa sa (35), pa ćemo najprije pokazati da se nepristranost procjenitelja S^2 može dokazati i bez dodatne pretpostavke 4.

Primjenjujući aparat linearne algebre najprije ćemo pokazati kako se zbroj $q = \sum_{i=1}^{n}[y_i - \mu_{\hat{\boldsymbol{a}}}(x_i)]^2$ iz (48) može transformirati u zbroj od $n-r$ kvadratnih pribrojnika.

Budući da su stupci matrice $\boldsymbol{X}$ linearno nezavisni vektori $\boldsymbol{x}^{(j)} \in \mathbf{R}^n$ $(j = 1, \ldots, r)$, postoji takva ortonormirana baza $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r, \ldots, \boldsymbol{e}_n$ u $\mathbf{R}^n$ da se svaki $\boldsymbol{x}^{(j)}$ može prikazati kao linearna kombinacija baznih vektora $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r$. Stoga se i vektor $\boldsymbol{\mu} = \sum_{j=1}^{r} a_j \boldsymbol{x}^{(j)}$ $(a_j \in \mathbf{R})$ može prikazati kao linearna kombinacija istih baznih vektora, tj. može se pisati

$$\text{(49)} \qquad \boldsymbol{\mu} = \sum_{j=1}^{r} \alpha_j \boldsymbol{e}_j, \qquad \alpha_j \in \mathbf{R}, \quad j = 1, \ldots, r.$$

Vektor $\boldsymbol{y} \in \mathbf{R}^n$ može se, također, prikazati kao linearna kombinacija baznih vektora $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ pa pišemo

(50) $$\boldsymbol{y} = \sum_{i=1}^{n} \beta_i \boldsymbol{e}_i , \qquad \beta_i \in \mathbf{R} , \quad i = 1, \ldots, n .$$

Primjenjujući oznaku $\|\cdot\|$ za euklidsku normu (apstraktnu duljinu) vektora, možemo pisati

(51) $$\|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = \|\sum_{j=1}^{r}(\beta_j - \alpha_j)\boldsymbol{e}_j + \sum_{i=r+1}^{n} \beta_i \boldsymbol{e}_i\|^2 = \sum_{j=1}^{r}(\beta_j - \alpha_j)^2 + \sum_{i=r+1}^{n} \beta_i^2 .$$

Imajući na umu (12) i postupak (MNK) kojim je dobiveno s^2 definirano u (48), očigledno je da iz (51) proizlazi

(52) $$q = (n-r)s^2 = \min_{(a_1,\ldots,a_r)} \|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = \min_{(\alpha_1,\ldots,\alpha_r)} \|\boldsymbol{y} - \boldsymbol{\mu}\|^2 = \sum_{i=r+1}^{n} \beta_i^2 ,$$

gdje je

(53) $$\beta_i = \boldsymbol{y}\, \boldsymbol{e}_i^{\mathsf{T}} , \qquad i = 1, \ldots, n .^*$$

Tretiramo li vektore $\boldsymbol{y}$ i $\boldsymbol{\mu}$ kao vrijednosti slučajnih vektora $\boldsymbol{Y}$ i $\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{X})$ iz r-dimenzionalnoga linearnog regresijskog modela, brojeve q i β_i tertirat ćemo kao vrijednosti odgovarajućih slučajnih varijabli Q i B_i, pa jednadžbi (52) odgovara pripadna jednadžba sa slučajnim varijablama

(54) $$Q = (n-r)S^2 = \sum_{i=r+1}^{n} B_i^2 ,$$

dok jednadžbi (53) odgovara

(55) $$B_i = \boldsymbol{Y}\, \boldsymbol{e}_i^{\mathsf{T}} , \qquad i = 1, \ldots, n .$$

Uzimajući u obzir modelsku jednadžbu (9), (55) postaje

$$B_i = (\boldsymbol{\mu} + \boldsymbol{\mathcal{E}})\, \boldsymbol{e}_i^{\mathsf{T}} = \boldsymbol{\mu}\, \boldsymbol{e}_i^{\mathsf{T}} + \boldsymbol{\mathcal{E}}\, \boldsymbol{e}_i^{\mathsf{T}} .$$

Iz (49) se vidi da je

$$\boldsymbol{\mu}\, \boldsymbol{e}_i^{\mathsf{T}} = 0 , \qquad \text{za } i = r+1, \ldots, n ,$$

tako da se može pisati

(56) $$B_i = \boldsymbol{\mathcal{E}}\, \boldsymbol{e}_i^{\mathsf{T}} , \qquad i = r+1, \ldots, n .$$

*Vektore $\boldsymbol{e}_i \in \mathbf{R}^n$ $(i = 1, \ldots, n)$ također shvaćamo kao jednoredne matrice i stoga se skalarni produkt piše kao $\boldsymbol{y}\boldsymbol{e}_i^{\mathsf{T}}$.

Promotri li se slučajni vektor $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, gdje je $Z_i = \boldsymbol{\mathcal{E}}\, \boldsymbol{e}_i^{\mathsf{T}}$ $(i = 1, \ldots, n)$, odmah se vidi da se može pisati

(57) $$\boldsymbol{Z} = \boldsymbol{\mathcal{E}}\, \boldsymbol{E} ,$$

gdje je $\boldsymbol{E}$ ortogonalna matrica n-tog reda ($\boldsymbol{E}^{\mathsf{T}} \boldsymbol{E} = \boldsymbol{I}_n$), čiji su stupci sastavljeni od komponenata vektora $\boldsymbol{e}_i$ u standardnoj bazi prostora $\mathbf{R}^n$.

Primijene li se na slučajni vektor $\boldsymbol{Z}$ formule iz zad. 23. u V. poglavlju, te uzme u obzir (10), dobiva se

(58) $$\begin{cases} \mathrm{E}[\boldsymbol{Z}] = \mathrm{E}[\boldsymbol{\mathcal{E}}]\, \boldsymbol{E} = \boldsymbol{O} , \\ \boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{E}^{\mathsf{T}} \boldsymbol{\Sigma}_{\boldsymbol{\mathcal{E}}} \boldsymbol{E} = \sigma^2 \boldsymbol{E}^{\mathsf{T}} \boldsymbol{E} = \sigma^2 \boldsymbol{I}_n , \end{cases}$$

a to znači da je

(59) $$\begin{cases} \mathrm{E}[Z_i] = \mathrm{E}[B_i] = 0 , & i = r+1, \ldots, n \\ \mathrm{V}[Z_i] = \mathrm{V}[B_i] = \sigma^2 = \mathrm{E}[B_i^2] , & i = r+1, \ldots, n . \end{cases}$$

Iz (54) i (59) dobiva se

$$\mathrm{E}[Q] = (n-r)\mathrm{E}[S^2] = \sum_{i=r+1}^{n} \mathrm{E}[B_i^2] = (n-r)\sigma^2 ,$$

iz čega neposredno slijedi da je $\mathrm{E}[S^2] = \sigma^2$, čime je, bez prizivanja na dodatnu pretpostavku 4, dokazana nepristranost procjenitelja S^2 za nepoznati parametar σ^2.

Relacija (57) pokazuje, nadalje, da je slučajni vektor $\boldsymbol{Z}$ dobiven ortogonalnom linearnom transformacijom slučajnog vektora $\boldsymbol{\mathcal{E}}$ kojemu pripada n-dimenzionalna normalna razdioba $\mathrm{N}(\boldsymbol{O}, \sigma^2 \boldsymbol{I}_n)$ s vektorom očekivanja $\boldsymbol{O}$ i kovarijancnom matricom $\sigma^2 \boldsymbol{I}_n$ (v. (21)). Stoga će slučajnom vektoru $\boldsymbol{Z}$ također pripadati n-dimenzionalna normalna radioba $\mathrm{N}(\boldsymbol{O}, \sigma^2 \boldsymbol{I}_n)$ (v. točku 10. u V.6), što znači da su $Z_1, \ldots, Z_n$ nezavisne slučajne varijable i vrijedi $Z_i \sim \mathrm{N}(0, \sigma^2)$ $(i = 1, \ldots, n)$.

Budući da je $B_i = Z_i$, za $i = r+1, \ldots, n$, očigledno je da vrijedi

(60) $$\frac{1}{\sigma} B_i \sim \mathrm{N}(0,1) , \qquad i = r+1, \ldots, n ,$$

pa iz (54), (60) i točke 5. u V.6. proizlazi

$$\frac{n-r}{\sigma^2} S^2 = \frac{1}{\sigma^2} Q = \sum_{i=r+1}^{n} \left(\frac{1}{\sigma} B_i\right)^2 \sim \chi^2(n-r) ,$$

čime je u potpunosti dokazan fundamentalni teorem.

Da bi se dokazala relacija (42) primijetimo da slučajnoj varijabli $\dfrac{\widehat{A}_j - a_j}{\sigma\sqrt{b_{jj}}}$ pripada standardna normalna razdioba $\mathrm{N}(0,1)$ i da se može pisati

$$T_j = \frac{\widehat{A}_j - a_j}{\sigma\sqrt{b_{jj}}} \sqrt{\frac{n-r}{\frac{n-r}{\sigma^2}S^2}},$$

pa primjenom fundamentalnog teorema i točke 7. iz V.6. odmah slijedi tvrdnja (42).

Primjenom fundamentalnog teorema dobiva se još jedan rezultat koji ima veliku praktičnu vrijednost, što ćemo vidjeti u daljnjim razmatranjima.

Uzme li se, naime, prirodni broj $m(1 \leq m \leq r)$ i linearno nezavisni vektori $x^{(k)} \in \mathbf{R}^n (k = 1, \ldots, r-m)$, koji se mogu prikazati kao linearne kombinacije baznih vektora $e_1, \ldots, e_{r-m}$, tada se i vektor $\widetilde{\mu} = \sum_{k=1}^{r-m} \widetilde{a}_k x^{(k)}$ $(\widetilde{a}_k \in \mathbf{R})$ može prikazati kao linearna kombinacija istih baznih vektora, tj. može se pisati

$$\widetilde{\mu} = \sum_{k=1}^{r-m} \widetilde{\alpha}_k e_k, \qquad \widetilde{\alpha}_k \in \mathbf{R}, \qquad k = 1, \ldots, r-m. \tag{61}$$

Iz (50) i (61) proizlazi da je

$$\|y - \widetilde{\mu}\|^2 = \sum_{k=1}^{r-m} (\beta_k - \widetilde{\alpha}_k)^2 + \sum_{i=r-m+1}^{n} \beta_i^2,$$

iz čega proizlazi da je

$$\widetilde{q} = \min_{(\widetilde{a}_1, \ldots, \widetilde{a}_{r-m})} \|y - \widetilde{\mu}\|^2 = \min_{(\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_{r-m})} \|y - \widetilde{\mu}\|^2 = \sum_{i=r-m+1}^{n} \beta_i^2. \tag{62}$$

Naposljetku, iz (52) i (62) dobivamo

$$\widetilde{q} - q = \sum_{i=r-m+1}^{r} \beta_i^2,$$

odnosno u terminima slučajnih varijabli

$$\widetilde{Q} - Q = \sum_{i=r-m+1}^{r} B_i^2, \tag{63}$$

pri čemu vektore y i μ, odnosno $\widetilde{\mu}$, razmatramo kao vrijednosti slučajnih vektora iz odgovarajućih modela višestruke linearne regresije.

Na temelju (60), (63) i rezultata navedenih u točki 5. i točki 7. iz V.6, odmah slijedi vrlo važna relacija

$$\frac{n-r}{m}\,\frac{\widetilde{Q} - Q}{Q} \sim \mathrm{F}(m, n-r), \tag{64}$$

koja će omogućiti rješavanje problema testiranja hipoteza o skupini regresijskih koeficijenata.

7. Testiranje hipoteza o skupini regresijskih koeficijenata

U regresijskoj analizi uobičajeno je da se ulazne varijable zovu faktori, tako da se, pojednostavnjeno govoreći, analiza zapravo sastoji od utvrđivanja utjecaja pojedinih faktora na promatranu izlaznu slučajnu varijablu. Stoga se često, umjesto o jednodimenzionalnome i višedimenzionalnome regresijskom modelu, govori o *jednofaktorskoj* i *višefaktorskoj (multifaktorskoj) regresijskoj analizi*. Glede toga nameće se pitanje kako utvrditi značajnost utjecaja pojedinog faktora ili skupine faktora na vjerojatnosnu razdiobu izlazne slučajne varijable. Preciznije govoreći, postavlja se zadatak da se, uz zadanu razinu značajnosti, testira hipoteza da određena skupina faktora nema utjecaja na izlaznu varijablu.

Pretpostavima da je riječ o r-dimenzionalnome $(r \geq 2)$ linearnom regresijskom modelu i postavimo nul-hipotezu

$$\mathrm{H}_0 : a_r = a_{r-1} = \ldots = a_{r-m+1} = 0 \qquad 1 \leq m < r,$$

tj. da je m regresijskih koeficijenata nula, odnosno da postoji samo $r-m$ faktora, među promatranih r faktora, koji značajno utječu na izlaznu slučajnu varijablu.

Alternativna je hipoteza da bar jedan od spomenutih m regresijskih koeficijenata nije jednak nuli.

Da bi se riješio postavljeni zadatak treba definirati prikladnu test-statistiku, za koju će se moći odrediti pripadna vjerojatnosna razdioba u uvjetima istinitosti hipoteze H_0. Do takve statistike može se doći ovim heurističkim razmatranjem: Ako regresijska funkcija ima oblik

$$\mu_{a}(x) = \sum_{j=1}^{r} a_j x^{(j)}, \tag{65}$$

onda veličina

$$\sum_{i=1}^{n} [y_i - \mu_{a}(x_i)]^2, \qquad a = (a_1, \ldots, a_r) \in \mathbf{R}^r,$$

pokazuje "kvalitetu" prilagodbe funkcije $x \mapsto \mu_{a}(x)$ danim podacima (x_i, y_i) $(i = 1, \ldots, n)$. Odrede li se procjene $\widehat{a}_j$ $(j = 1, \ldots, r)$ regresijskih koeficijenata a_j po metodi najmanjih kvadrata, dobiva se vrijednost

$$q = \sum_{i=1}^{n} [y_i - \mu_{\widehat{a}}(x_i)]^2. \tag{66}$$

Ako se, pak, pretpostavi da regresijska funkcija sadrži samo $r-m$ koeficijenata (različitih od nule), tj. da ima oblik

(67) $$\mu_{\widetilde{a}}(x) = \sum_{k=1}^{r-m} \widetilde{a}_k x^{(k)} ,$$

i da su $\widetilde{a}_k$ $(k = 1, \ldots, r-m)$ također dobiveni kao MNK-procjene odgovarajućih nepoznatih regresijskih koeficijenata, stavit će se

(68) $$\widetilde{q} = \sum_{i=1}^{n} [y_i - \mu_{\widetilde{a}}(x_i)]^2 , \qquad \widetilde{a} = (a_1, \ldots, a_{r-m}) \in \mathbf{R}^{r-m} .$$

Očigledno je $q \leq \widetilde{q}$, jer je q dobiveno kao najmanja vrijednost zbroja kvadrata odstupanja empirijskih izlaznih vrijednosti y_i $(i = 1, \ldots, n)$ od odgovarajućih teorijskih vrijednosti najbolje prilagođene regresijske funkcije sa r slobodnih parametara, dok je $\widetilde{q}$ dobiveno na isti način, ali sa samo $r - m$ slobodnih parametara, što ne može dati manju vrijednost zbroja odgovarajućih kvadrata odstupanja.

Možemo, stoga, konstatirati da razlika $\widetilde{q} - q$ pokazuje koliko se pogoršala prilagodba regresijske funkcije danim podacima, kada se smanji za m broj regresijskih koeficijenata u r-dimenzionalnome linearnom regresijskom modelu. Očigledno je da sama vrijednost $\widetilde{q} - q$ neće moći poslužiti kao kriterij za donošenje odluke o prihvaćanju, odnosno odbacivanju hipoteze H_0, jer je riječ o veličini koja ovisi o upotrijebljenim mjernim jedinicama, pa je prirodno da se potraži test-statistika čije vrijednosti su "čisti" brojevi.

Na temelju (47) i (66) može se reći da je veličina $\dfrac{q}{n-r}$ dobra procjena za nepoznati parametar σ^2, neovisno o hipotezi H_0, pa se može pisati

(69) $$q \approx (n-r)\sigma^2 .$$

Ako je hipoteza H_0 stvarno istinita, onda se zbog istih razloga može smatrati da je i veličina $\dfrac{\widetilde{q}}{n-r+m}$ dobra procjena za σ^2, što opravdava zapis

(70) $$\widetilde{q} \approx (n-r+m)\sigma^2 .$$

Iz (69) i (70) proizlazi

$$\widetilde{q} - q \approx m\sigma^2 ,$$

što opravdava zaključak da je i veličina $\dfrac{\widetilde{q}-q}{m}$ dobra procjena za σ^2.

Ako, pak, hipoteza H_0 nije istinita, onda se može očekivati da će veličina $\dfrac{\widetilde{q}-q}{m}$ biti značajno veća od σ^2, jer se reduciranjem broja regresijskih koeficijenata značajno pogoršala prilagodba regresijske funkcije danim empirijskim podacima. Stoga se čini razumnim smatrati da bi vrijednost omjera

(71) $$v = \frac{\dfrac{\widetilde{q}-q}{m}}{\dfrac{q}{n-r}} = \frac{n-r}{m} \, \frac{\widetilde{q}-q}{q}$$

mogla poslužiti kao kriterij za donošenje odluke o hipotezi H_0.

To postaje naročito prihvatljivo kada se v tretira kao vrijednost slučajne varijable

(72) $$V = \frac{n-r}{m} \, \frac{\widetilde{Q}-Q}{Q} ,$$

za koju se odmah razabire (v. (64)) da joj pripada F-razdioba sa $(m, n-r)$ stupnjeva slobode.

Prema tome, prevelika vrijednost v, test-statistike V, sugerirat će da hipotezu H_0 treba odbaciti. Kritično područje zadane razine značajnosti α odredit će se uvjetom

(73) $$v \geq F_{m,n-r}^{-1}(1-\alpha) ,$$

gdje je $F_{m,n-r}^{-1}$ inverzna funkcija za f.r.v. F-razdiobe sa $(m, n-r)$ stupnjeva slobode, čije su vrijednosti prikazane u tabl. VII. u Dodatku.

Ilustrirajmo opisani postupak na 1. primjeru, tako da postavimo hipotezu $H_0 : a_1 = a_2 = 0$, što nam donekle i sugeriraju vrijednosti procjena regresijskih koeficijenata navedene u tabl. 4. Praktički govoreći, postavljamo hipotezu da količina cementa i agregata ne utječu na tlačnu čvrstoću betonskih kocki.

Imamo, dakle, $n = 7$, $r = 4$, $m = 2$ i $q = (n-r)s^2 = 3 \cdot 0{,}72 = 2{,}16$, pa ostaje da se još izračuna

$$\widetilde{q} = \min_{a_3, a_4} \sum_{i=1}^{7} [y_i - a_3 x_i^{(3)} - a_4]^2 = \sum_{i=1}^{7} [y_i - \widetilde{a}_3 x_i^{(3)} - \widetilde{a}_4]^2 = 2{,}89 ,$$

pri čemu je $\widetilde{a}_3 \approx -73{,}9$ i $\widetilde{a}_4 \approx 72{,}5$.

Sada se, prema (71), može izračunati

$$v = \frac{7-4}{2} \cdot \frac{2{,}89 - 2{,}16}{2{,}16} \approx 0{,}51 .$$

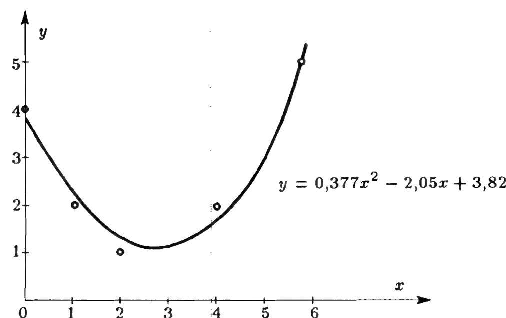
Uzme li se $\alpha = 0{,}05$, iz tabl. VII. u Dodatku odčitava se $F_{2,3}^{-1}(0{,}95) = 9{,}55$, pa se vidi da vrijednost (0,51) test-statistike ostaje izvan kritičnog područja $[9{,}55 ; \infty\rangle$, što znači da hipotezu H_0 treba prihvatiti.

8. Nelinearna regresija

U mnogim praktičnim situacijama priroda promatranog fenomena sugerirat će nam da za regresijsku funkciju ne valja uzeti linearnu, odnosno afinu funkciju. Ako je riječ o jednodimenzionalnoj regresiji, onda će nam to redovito pokazati i grafički prikaz podataka (x_i, y_i) $(i = 1, \ldots, n)$ u pravokutnom koordinatnom sustavu. Tako, na primjer, dobiju li se mjerenjem podaci navedeni u tabl. 7, čiji je grafički prikaz dan na sl. 35, očigledno je da će trebati uzeti neku nelinearnu funkciju regresije da bi se dobila dobra prilagodba danim podacima. Prva je pomisao da se uzme polinom drugog stupnja (kvadratna funkcija) kao regresijska funkcija.

Tablica 7.

x_i	0	1	2	4	6
y_i	4	2	1	2	5



Slika 35. Grafički prikaz podataka iz tabl. 7.

Budući da nema posebnih teškoća, odmah ćemo razmotriti opći problem tzv. ***polinomske regresije***, u kojem se pretpostavlja da regresijska funkcija ima oblik polinoma $(r-1)$-og stupnja

(74) $$\mu_{\boldsymbol{a}}(x) = a_1 x^{r-1} + a_2 x^{r-2} + \ldots + a_{r-1}x + a_r \,.$$

Zadatak nam je odrediti procjene $\widehat{a}_j$ $(j = 1,\ldots,r)$ koeficijenata a_j, tako da vrijedi

(75) $$\min_{(a_1,\ldots,a_r)} \sum_{i=1}^{n}[y_i - \mu_{\boldsymbol{a}}(x_i)]^2 = \sum_{i=1}^{n}[y_i - \mu_{\widehat{\boldsymbol{a}}}(x_i)]^2 \,,$$

gdje je $\widehat{\boldsymbol{a}} = (a_1,\ldots,a_r)$ MNK-procjena vektorskog parametra $\boldsymbol{a} = (a_1,\ldots,a_r)$.

Usporedbom (74) sa (5) odmah se vidi da se model polinomske regresije može shvatiti i kao model r-dimenzionalne linearne regresije u kojem je $x^{(j)} = x^{r-j}$ $(j = 1,\ldots,r)$, tako da se i svi rezultati izvedeni u r-dimenzionalnome linearnom regresijskom modelu mogu prenijeti na odgovarajući model polinomske regresije, pri čemu se mora paziti da se u konkretnim računima $x_i^{(j)}$ zamjenjuje sa x_i^{r-j} $(i = 1,\ldots,n)$.

Slika 35. sugerira nam da podacima iz tabl. 7. pokušamo prilagoditi kvadratni polinom $(r = 3)$, kao regresijsku funkciju. Da bismo izračunali MNK-procjene odgovarajućih koeficijenata načinimo tablicu 8.

Tablica 8.

x_i^2	x_i^1	x_i^0	y_i
0	0	1	4
1	1	1	2
4	2	1	1
16	4	1	2
36	6	1	5

Sada je vidljivo da je matrica ulaznih podataka

$$\boldsymbol{X} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 16 & 4 & 1 \\ 36 & 6 & 1 \end{bmatrix},$$

dok je vrijednost vektora izlaznih podataka

$$y = (4, 2, 1, 2, 5)\,,$$

pa se primjenom formule (16) dobiva $\widehat{a}_1 = 0{,}377$, $\widehat{a}_2 = -2{,}05$, $\widehat{a}_3 = 3{,}82$.

Može se, dakle, ustanoviti da je podacima iz tabl. 7. prilagođen kvadratni polinom

$$\mu(x) = 0{,}377\,x^2 - 2{,}05\,x + 3{,}82\,,$$

kao pripadna regresijska funkcija.

Na temelju formule (34) može se izračunati i procjena s^2 za nepoznati parametar σ^2, što omogućuje (formule (40), (43) i (46)) i izradbu tablice 9, koja je analogna tablicama 4, 5. i 6.

Iz tabl. 9. mogu se očitavati intervali povjerenja pouzdanosti $\gamma = 0{,}95$ za regresijske koeficijente, te vrijednosti odgovarajućih test-statistika za testiranje hipoteze da je regresijski koeficijent jednak nuli. Budući da je u promatranom

Tablica 9.

j	1	2	3
a_j	0,377	−2,05	3,82
$\widehat{\sigma}_j$	0,0426	0,268	0,304
$g_1^{(j)}$	0,194	−3,20	2,52
$g_2^{(j)}$	0,561	−0,90	5,13
t_j	8,86	−7,65	12,6

primjeru $n = 5$ i $r = 3$, za razinu značajnosti $\alpha = 0{,}10$, dobiva se

$$G_{n-r}^{-1}\left(1-\frac{\alpha}{2}\right) = G_2^{-1}(0{,}95) = 5{,}99$$

(v. tabl. VI. u Dodatku), pa se zaključuje da svaku od hipoteza $H_{0j} : a_j = 0$ $(j = 1, 2, 3)$ treba odbaciti. To praktički znači da se ne može proći s manje od tri nenulta koeficijenta u regresijskoj funkciji polinomskog tipa.

Pripadna tablica analize varijance izgleda ovako:

Tablica 10.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Srednje kvadratno odstupanje	Koeficijent determinacije (korigirani)
model	2	10,57	2,11	
slučajna greška	2	0,23	0,05	
ukupno rasipanje	4	10,80		0,978 (0,956)

Može se, dakle, konstatirati da se 97,8 %, odnosno 95,6 %, ukupnog rasipanja može protumačiti modelom, što upućuje na zaključak da je model dobro prilagođen danim empirijskim podacima. Da se, kao regresijska funkcija, uzela afina funkcija ($\mu(x) = ax + b$), dobio bi se koeficijent determinacije $R^2 = 0,125$, što znači da bi se modelom moglo objasniti samo 12,5 % ukupnog rasipanja.

Rezultati izvedeni na temelju modela višestruke linearne regresije mogu se primijeniti i na općenitiji slučaj nelinearne regresije, gdje se pretpostavlja da regresijska funkcija ima oblik

(76) $$\mu_{\boldsymbol{a}}(x) = a_1 h_1(x) + a_2 h_2(x) + \ldots + a_r h_r(x)\,,$$

pri čemu su $h_1, \ldots, h_r$ određene realne funkcije. Usporedi li se (76) sa (5), vidi se da će se te formule podudarati kada se stavi $x^{(j)} = h_j(x)$ $(j = 1, \ldots, r)$, što znači da se MNK-procjene $\widehat{a}_j$, za nepoznate parametre a_j u (76), mogu dobiti primjenom formule (16), imajući na umu da je sada $x_i^{(j)} = h_j(x_i)$ $(i = 1, \ldots, n)$.

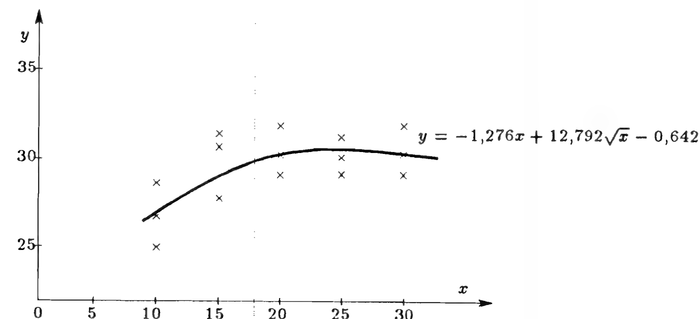
Za ilustraciju opisane problematike razmotrimo idući primjer.

2. primjer

Da bi se ustanovila ovisnost tlačne čvrstoće betona Y (MPa) o koncentraciji x (promila) određenog aditiva izveden je eksperiment, a dobiveni podaci prikazani su u tabl. 11. Uz svaku navedenu koncentraciju (x_i) aditiva izrađene su po tri betonske kocke na kojima je mjerena tlačna čvrstoća (y_i).

Tablica 11.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i		10			15			20			25			30	
y_i	25	27	29	28	30	31	30	31	33	30	31	32	30	31	33



Slika 36. Grafički prikaz podataka iz tabl. 11.

Pretpostavimo da je regresijska funkcija oblika

(77) $$\mu_{\boldsymbol{a}}(x) = a_1 x + a_2\sqrt{x} + a_3\,,$$

što s obzirom na relaciju (76) znači da smo uzeli $h_1(x) = x$, $h_2(x) = \sqrt{x}$ i $h_3(x) = 1$.

Osnova za sve daljnje proračune je tabl. 12.

Tablica 12.

x_i	$\sqrt{x_i}$	x_i^0	y_i
10	3,16	1	25
10	3,16	1	27
10	3,16	1	29
15	3,87	1	28
15	3,87	1	30
15	3,87	1	31
20	4,47	1	30
20	4,47	1	31
20	4,47	1	33
25	5	1	30
25	5	1	31
25	5	1	32
30	5,48	1	30
30	5,48	1	31
30	5,48	1	33

Rezultati su prikazani u tabl. 13.

Tablica 13.

j	1	2	3
$\widehat{a}_j$	−1,276	12,792	−0,642
$\widehat{\sigma}_j$	0,6736	5,8191	12,201
$g_1^{(j)}$	−2,744	0,1097	−27,234
$g_2^{(j)}$	0,192	25,4736	25,950
t_j	−1,895	2,1982	−0,053

Procjena regresijske funkcije oblika (77) za dane empirijske podatke, prema tome, glasi

$$\mu(x) = -1,276\,x + 12,792\sqrt{x} - 0,642\,.$$

Naravno da smo, umjesto funkcije regresije (77), mogli uzeti i neki drugi tip regresijske funkcije (v. zad. 14), pa se odmah otvara problem egzaktnog vrednovanja pojedinoga regresijskog modela sa stajališta njegove prilagodbe danim podacima, o čemu će biti više riječi u XIII.5.

Izbor tipa regresijske funkcije, teorijski gledano, proizvoljan je postupak. Međutim, istraživač se pri takvom izboru ipak oslanja na neke spoznaje o promatranim veličinama, ili ga pak grafički prikaz podataka navodi na ideju o tipu regresijske funkcije. Svakako je najočiglednija situacija kada se pokaže da točke (x_i, y_i) $(i = 1, \ldots, n)$, bar približno, leže na jednom pravcu. Ako to nije tako, nastoje se originalni podaci (x_i, y_i) transformirati tako, ako je to moguće, da transformirani podaci (x'_i, y'_i) približno leže na jednom pravcu. Takvo se rezoniranje temelji, primjerice, na činjenici da eksponencijalna regresijska zavisnost $\mu(x) = b\exp(ax)$ između ulazne (x) i izlazne (y) varijable uzrokuje afinu regresijsku zavisnost između varijabli $x' = x$ i $y' = \ln y$. Iz $y = b\exp(ax)$ proizlazi, naime, da je $\ln y = \ln b + ax$, odnosno $y' = a'x' + b'$, gdje je $a' = a$ i $b' = \ln b$.

Ako se, prema tome, utvrdi afina regresijska zavisnost na podacima (x'_i, y'_i), tj. odrede se MNK-procjene $\widehat{a}'$ i $\widehat{b}'$ odgovarajućih parametara a' i b' afine funkcije $x' \mapsto a'x' + b'$, može se smatrati da između izvornih podataka (x_i, y_i) postoji *eksponencijalna regresijska zavisnost* za koju su $\widehat{a} = \widehat{a}'$ i $\widehat{b} = \exp(\widehat{b}')$ procjene odgovarajućih parametara.

Sluti li se, recimo, da bi se moglo raditi o regresijskoj ovisnosti tipa potencije, tj. da regresijska funkcija ima oblik $\mu(x) = bx^a$, izvest će se transformacija izvornih podataka formulama $x' = \ln x$ i $y' = \ln y$, jer će tada x' i y' biti povezani afinom funkcijom $x' \mapsto a'x' + b'$, gdje je $a = a'$ i $b = \exp(b')$.

Numerički postupak opet se provodi na transformiranim podacima $(x'_i, y'_i) = (\ln x_i, \ln y_i)$ $(i = 1, \ldots, n)$, kojim se dolazi do MNK-procjena $\widehat{a}'$ i $\widehat{b}'$, nepoznatih parametara a' i b' afine funkcije, a kojima odgovaraju procjene $\widehat{a} = \widehat{a}'$ i $\widehat{b} = \exp(\widehat{b}')$ parametara a i b *potencijske regresijske funkcije* $\mu(x) = bx^a$.

Primjedba

Iz prikazane teorije regresijske analize, posebno višestruke regresije, te razmotrenih primjera, očigledno je da dobivanje traženih rezultata redovito zahtijeva vrlo opsežna numerička računanja. Primjena računala i odgovarajućih posebnih programa za rješavanje regresijskih problema omogućuje da se brzo i lako rješavaju i oni praktični problemi koji sadrže velik broj podataka i koji bi bez upotrebe računala bili praktički nerješivi. Osim toga, primjena računala omogućuje da se na iste podatke primijene različiti regresijski modeli i brzo dobije odgovor na pitanje o većoj ili manjoj prikladnosti pojedinog modela. Isto tako omogućena je i velika raznolikost u prikazivanju rezultata, pa se nakon računalske obrade problema mogu dobiti različite tablice (tablica analize varijance i dr.) i pripadni grafički prikazi (krivulje regresije, intervali povjerenja i sl.).

Zadaci

1. Provjerite da se linearni regresijski model opisan u XI.1. i XI.2. dobiva kao specijalni slučaj r-dimenzionalnoga linearnog regresijskog modela za $r = 2$ i $\boldsymbol{a} = (a, b)$.

2. Izvedite formule (4) i (5) iz XI.1. primjenom formule (16).

3. Pokažite da se formule (11) i (12) iz XI.1. mogu dobiti kao posebni slučaj formula (20) i (22).

4. Pokažite da se sustav jednadžbi (17) može matrično zapisati u obliku (15), gdje umjesto $\widehat{\boldsymbol{a}}$ stoji $\boldsymbol{a}$.

5. Neka je $(X_1, \ldots, X_n)$ slučajni uzorak za slučajnu varijablu X, kojoj pripada očekivanje μ i varijanca σ^2. Neka je $\widehat{T} = \sum_{i=1}^{n} \alpha_i X_i$ $(\alpha_i \in \mathbf{R})$ određena statistika (linearna statistika).

a) Nađite dovoljan uvjet da $\widehat{T}$ bude NL-procjenitelj za μ.

b) Dokažite da je aritmetička sredina slučajnog uzorka $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ najbolji linearni procjenitelj (NLN-procjenitelj) za parametar μ u smislu da je $\mathrm{V}[\overline{X}] \leq \mathrm{V}[\widehat{T}]$, za svaki NL-procjenitelj $\widehat{T}$.

6. Dokažite da u r-dimenzionalnome linearnom regresijskom modelu vrijedi $s_y^2 = \widehat{\sigma}_0^2 + \widehat{\sigma}^2$ (formula (37)).

7. Načinite tablicu analize varijance na temelju podataka iz 1. primjera.

8. Izvedite formule (41).

Uputa: Iskoristite činjenicu da $\widehat{A}_j$ približno ima normalnu razdiobu $\mathrm{N}(a_j, s^2 b_{jj})$.

9. Izvedite formule za granice intervala povjerenja zadane pouzdanosti γ za nepoznatu vrijednost regresijske funkcije $\mu_{\boldsymbol{a}}(\boldsymbol{x})$, uz pretpostavku da je broj n podataka vrlo velik.

Uputa: Primijenite asimptotsku normalnost slučajne varijable $\mu_{\widehat{\boldsymbol{A}}}(\boldsymbol{x})$.

10. Pokažite da se formule (29) i (30) iz XI.2. mogu dobiti kao posebni slučaj $(r = 2)$ iz formula (42).

11. Pokažite da se formula (56) iz XI.4. može dobiti kao posebni slučaj formule (72) za $r = 2$ i $m = 1$.

12. Izvedite relacije (44), (45) i (46) kojima se definiraju kritična područja pri testiranju jednostavne hipoteze $\mathrm{H}_0 : a_j = a_0$, prema odgovarajućoj alternativnoj hipotezi $\mathrm{H}_1 : a_j \neq a_0$ (ili $a_j < a_0$, ili $a_j > a_0$).

13. Napravite proračun svih relevantnih veličina na temelju podataka iz 2. primjera uz pretpostavku da je regresijska funkcija polinom trećeg stupnja.

14. Na temelju podataka iz 2. primjera izračunajte procjene nepoznatih parametara uz pretpostavku da je regresijska funkcija:

(a) eksponencijalnog tipa,

(b) potencijskog tipa.

Nacrtajte i odgovarajuće grafičke prikaze.

15. Želi se istražiti ovisnost između početne težine tovljenika ($x^{(1)}$), količine pojedene hrane ($x^{(2)}$) i konačne težine (y). Eksperimentom su dobivene ove vrijednosti:

$x_i^{(1)}$	$x_i^{(2)}$	y_i
32	173	50
33	226	77
33	259	80
36	183	70
39	311	97
38	235	84
40	230	92
41	236	80
42	272	95
45	292	100

a) Primjenom modela višestruke linearne regresije izračunajte pripadne regresijske koeficijente i napišite odgovarajuću procjenu za regresijsku funkciju.

b) Načinite tablicu analize varijance.

c) Odredite intervale povjerenja pouzdanosti $\gamma = 0{,}95$ za regresijske koeficijente.

d) Testirajte nul-hipotezu da je koeficijent uz varijablu $x^{(1)}$ jednak nuli.

16. Mjerena je učinkovitost (y) radnika svakog sata (x) tijekom radnog dana. Dobiveni su ovi rezultati:

Sat	1	2	3	4	5	6
Učinkovitost	150	148	175	165	172	155

a) Primjenom modela kvadratne regresije izračunajte pripadne regresijske koeficijente i napišite odgovarajuću procjenu za regresijsku funkciju.

b) Načinite tablicu analize varijance.

c) Usporedite dobivene rezultate s rezultatima dobivenim uz primjenu modela linearne regresije.

XIII. Analiza varijance

1. Uvod u problematiku

Analizu varijance, kao određeni matematički model i praktičnu tehniku za istraživanje nekih bioloških fenomena, prvi je razvio poznati engleski statističar R. A. Fisher (1890–1962). Danas je analiza varijance, koja se skraćeno zove ANOVA (prema engleskom: Analysis of Variance), vrlo važna i popularna metoda za istraživanje različitih slučajnih pojava u mnogim znanstvenim područjima.

U sklopu analize varijance razvijeno je nekoliko matematičkih modela (jednofaktorski, dvofaktorski i sl.), koji omogućuju operacionalizaciju vrlo jednostavnih postupaka za rješavanje važnih praktičnih zadataka.

Da bi se lakše i jasnije shvatila problematika i kasnije teorijske konstrukcije modela analize varijance, najprije će se razmotriti jedan vrlo tipičan primjer.

1. primjer

Tri različite tvornice automobila A, B i C proizvode, među ostalim, i tip automobila približno iste snage motora, pa se želi provjeriti hipoteza da potrošnja goriva ne ovisi o marki (tvornici) automobila. Kako organizirati eksperiment koji će omogućiti donošenje odluke o prihvaćanju, odnosno odbacivanju postavljene hipoteze?

Odmah se nameće ideja da se uzme nekoliko automobila svake marke, proveze određeni broj kilometara sa svakim od njih, te izmjeri potrošnja goriva. No, svaki vozač zna da potrošnja goriva ovisi i o mnogim drugim faktorima (vrsta ceste, vozačko iskustvo, godišnje doba i sl.). Da bi se eliminirao utjecaj ceste, mogu se svi automobili voziti po istoj cesti. Želi li se eliminirati i utjecaj vozača, čini se razumnim slučajno izabrati vozače, tako da se dobivene vrijednosti potrošnje goriva mogu smatrati vrijednostima slučajnog uzorka. U svakom slučaju cilj nam je utvrditi utjecaj samo jednog faktora, faktora proizvođača, na potrošnju goriva.

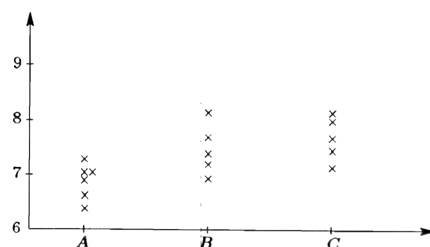
Recimo da je uzeto $n_1 = 6$ automobila marke A, $n_2 = 5$ automobila marke B i $n_3 = 5$ automobila marke C. Rezultati eksperimenta navedeni su u tabl. 1.

Tablica 1.

Marka automobila	Potrošnja goriva u litrama na 100 km						Aritmetička sredina
A	7,4	6,6	6,8	7,1	7,0	7,0	7,0
B	8,1	7,7	7,0	7,3	7,4		7,5
C	7,3	8,1	7,5	8,0	7,1		7,6

Može li se na temelju podataka iz tabl. 1. zaključiti da nema značajne razlike u potrošnji goriva među promatrane tri marke automobila?

Radi još veće jasnoće problema i upućivanja na vezu ovog zadatka s problemima regresijske analize, načinimo i određeni grafički prikaz podataka iz tabl. 1 (v. sl. 37).



Slika 37. Grafički prikaz podataka iz tabl. 1.

Budući da je marka automobila nenumeričko obilježje, na sl. 37. nije riječ o pravom koordinatnom sustavu, ali nam ona ipak sugerira da se izmjerene brojčane vrijednosti (litara/100 km) mogu shvatiti kao vrijednosti izlazne varijable, uzrokovane odgovarajućim vrijednostima (A,B,C) nenumeričkog faktora – marke automobila. To nam pokazuje da će matematički model za opisivanje promatranog fenomena imati određene sličnosti s regresijskim modelima, u smislu da se izlazna numerička vrijednost shvaća kao posljedica djelovanja nenumeričke vrijednosti (*razine*) ulazne varijable (*djelujućeg faktora*), u ovom slučaju marke automobila, koja određuje srednju vrijednost izlazne varijable (potrošnja goriva) i kojoj se dodaje slučajna greška.

Neka je, dakle, razina djelujućeg faktora A i pripadna srednja vrijednost izlazne varijable μ_A, tada se vrijednost izlazne varijable, recimo 7,4, shvaća kao zbroj $\mu_A + \varepsilon_A$, gdje je ε_A vrijednost slučajne varijable $\mathcal{E}_A$ (slučajna greška).

Prema tome, u opisanom modelu imat ćemo tri nepoznata parametra μ_A, μ_B i μ_C i još nepoznate parametre povezane sa slučajnim varijablama $\mathcal{E}_A$, $\mathcal{E}_B$ i $\mathcal{E}_C$. Pretpostavi li se, kao što se to obično čini, da su $\mathcal{E}_A$, $\mathcal{E}_B$ i $\mathcal{E}_C$ nezavisne slučajne varijable sa zajedničkom normalnom razdiobom $N(0,\sigma^2)$, onda je riječ samo još o nepoznatom parametru σ^2.

Brojeve iz prvog retka tabl. 1. možemo opisati jednadžbom

$$y_{iA} = \mu_A + \varepsilon_{iA}\,, \qquad i = 1,\ldots,6\,,$$

brojeve iz drugog retka jednadžbom

$$y_{iB} = \mu_B + \varepsilon_{iB}\,, \qquad i = 1,\ldots,5\,,$$

a brojeve iz trećeg retka jednadžbom

$$y_{iC} = \mu_C + \varepsilon_{iC}\,, \qquad i = 1,\ldots,5\,.$$

Sada se mogu formulirati različite statističke zadaće da se, na temelju danih podataka, utvrde određene činjenice o nepoznatim parametrima (točkasta i intervalna procjena parametara, testiranje različitih hipoteza i sl.).

Čini se prilično logičnim da se na početku postavljeno pitanje o postojanju ili nepostojanju značajne razlike u potrošnji goriva formulira kao problem testiranja nul-hipoteze $H_0 : \mu_A = \mu_B = \mu_C$, prema alternativnoj hipotezi da bar na jednom mjestu stoji znak nejednakosti.

To je tipični problem analize varijance i cijela teorija analize varijance uglavnom se sastoji od objašnjenja postupaka za njegovo rješavanje. Budući da se ta teorija uglavnom bavi analizom rasipanja (varijance) izlaznih podataka, teorija je i dobila naziv ***analiza varijance***.

Očigledno je da rasipanje aritmetičkih sredina (zadnji stupac tabl. 1) može poslužiti kao indikator valjanosti hipoteze H_0. Da smo, recimo, dobili sve tri aritmetičke sredine međusobno jednake, onda bi njihova varijanca (rasipanje) bila nula i u tom slučaju bismo smatrali da treba prihvatiti hipotezu H_0. U danom slučaju ta varijanca približno iznosi $s_1^2 = 0{,}58$, pa se postavlja pitanje da li je to dovoljno veliko za odbacivanje hipoteze H_0.

Pogledajmo kakvi se empirijski podaci mogu očekivati kada je hipoteza H_0 stvarno neistinita, tj. kada automobil marke A ima zaista manju potrošnju goriva od automobila marke B. Moglo se, recimo, dogoditi da je kod svih šest automobila marke A bila potrošnja 7 (litara/100 km), kod svih pet automobila marke B 7,5 (litara/100 km) i kod svih automobila marke C 7,6 (litara/100 km). Tada bi varijanca podataka unutar svakog retka tabl. 1. bila nula. U danom primjeru te varijance iznose $s_A^2 = 0{,}076$, $s_B^2 = 0{,}175$ i $s_C^2 = 0{,}190$, pa se možemo pitati da li dobivene vrijednosti upućuju na odbacivanje hipoteze H_0. Ponderirana sredina tih disperzija iznosi $s_2^2 = 0{,}14$ i u nastavku će se pokazati da je omjer $\frac{s_1^2}{s_2^2}$ prikladan indikator za donošenje odluke o hipotezi H_0, jer je očigledno da prevelika vrijednost toga omjera (veliko rasipanje aritmetičkih sredina redaka i malo rasipanje unutar redaka tabl. 1) upućuje na odbacivanje nul-hipoteze.

2. Jednofaktorski model

U 1. primjeru istaknuli smo bitne momente, koji će nam olakšati shvaćanje općih apstraktnih pojmova koje ćemo sada definirati kao ***jednofaktorski model analize varijance***.

Pretpostavlja se da je dano m ($m \geq 2$) nizova podataka

$$(1) \qquad \left\{ \begin{array}{ccc} y_{11}, & \cdots, & y_{1n_1} \\ y_{21}, & \cdots, & y_{2n_2} \\ \cdots & \cdots & \cdots \\ y_{m1}, & \cdots, & y_{mn_m} \end{array} \right. , \qquad n_1,\ldots,n_m \in \mathbf{N}\,,$$

i da je i-ti ($i = 1,\ldots,m$) niz dobiven mjerenjem slučajne varijable $Y_i \sim N(\mu_i,\sigma^2)$, te da su $Y_1,\ldots,Y_m$ nezavisne slučajne varijable. To znači da se y_{ij} ($i = 1,\ldots,m$, $j = 1,\ldots,n_i$) može interpretirati kao vrijednost slučajne varijable

$$(2) \qquad Y_{ij} = \mu_i + \mathcal{E}_{ij}\,,$$

gdje su $\mathcal{E}_{ij} \sim N(0,\sigma^2)$ nezavisne slučajne varijable.

Možemo, dakle, reći da je Y_{ij} izlazna slučajna varijabla, čije vrijednosti y_{ij} nastaju djelovanjem i-te razine (μ_i) određenog faktora, uz dodatak slučajne greške ($\mathcal{E}_{ij}$). Djelujući faktor (ulazna varijabla) najčešće ima nenumeričko obilježje i u tome je glavna razlika u odnosu na model jednodimenzionalne (jednofaktorske) regresijske analize.

U 1. primjeru djelujući je faktor marka automobila i imamo tri ($m = 3$) razine A, B i C toga faktora. U tabl. 1. navedena su mjerenja izlazne varijable y_{ij} (potrošnja goriva), pa za $i = 1$ imamo šest ($n_1 = 6$) vrijednosti, dok za $i = 2$ i $i = 3$ imamo pet ($n_2 = n_3 = 5$) vrijednosti izlazne varijable. Ukupno se raspolaže sa $n = n_1 + n_2 + n_3 = 16$ podataka o potrošnji goriva.

Općenito se stavlja

(3) $$n = \sum_{i=1}^{m} n_i ,$$

pri čemu n označuje ukupni broj podataka.

Sada možemo općenito definirati i glavni problem jednofaktorske analize varijance, koji se sastoji u određivanju postupka za testiranje nul-hipoteze

(4) $$\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_m ,$$

prema alternativnoj hipotezi da bar jedna jednakost u (4) nije istinita. Drugim riječima, problem se sastoji u određivanju kritičnog područja, zadane razine značajnosti, pri testiranju hipoteze o jednakosti očekivanja m nezavisnih slučajnih varijabli normalnih razdioba zajedničke nepoznate varijance σ^2, na temelju m nizova podataka (1).

U praktičnim situacijama hipoteza H_0 obično se iskazuje kao hipoteza da različite razine djelujućeg faktora ne utječu na promatranu izlaznu veličinu, odnosno da uočeni faktor nema utjecaj na promatranu veličinu.

Stavimo

(5) $$\mu = \frac{1}{n} \sum_{i=1}^{m} n_i \mu_i , \qquad \delta_i = \mu_i - \mu , \qquad i = 1, \ldots, m ,$$

pa je uobičajeno da se veličina μ zove *opća srednja vrijednost*, dok se δ_i zove *efekt* i-*te razine* djelujućeg faktora. U tom svjetlu modelska jednadžba (2) može se zapisati u obliku

(6) $$Y_{ij} = \mu + \delta_i + \mathcal{E}_{ij} , \qquad i = 1, \ldots, m , \quad j = 1, \ldots, n_i ,$$

a hipoteza H_0 iz (4) kao

(7) $$\mathrm{H}_0 : \delta_1 = \delta_2 = \ldots = \delta_m = 0 .$$

Modelska jednadžba (6) može se protumačiti tako da se izlazna vrijednost y_{ij} shvati kao zbroj opće srednje vrijednosti μ, efekta δ_i i-te razine djelujućeg faktora i vrijednosti ε_{ij} slučajne greške $\mathcal{E}_{ij}$. Hipotezom H_0, zapisanom u obliku (7), postavlja se teza da su efekti beznačajni.

Da bi se definirala prikladna test-statistika, pomoću koje će se odrediti kritično područje zadane razine značajnosti α, uvest će se najprije oznake

(8) $$\overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} , \qquad i = 1, \ldots, m ,$$

(9) $$\overline{Y} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n} \sum_{i=1}^{m} n_i \overline{Y}_i ,$$

(10) $$Q = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2 ,$$

(11) $$Q_1 = \sum_{i=1}^{m} n_i (\overline{Y}_i - \overline{Y})^2 ,$$

(12) $$Q_2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 .$$

Lako se provjerava da je (v. zad. 2)

(13) $$Q = Q_1 + Q_2 .$$

Iz (8) se vidi da je $\overline{Y}_i$ aritmetička sredina i-tog niza podataka, tj. onih izlaznih varijabli na koje djeluje i-ta razina promatranog faktora, dok je $\overline{Y}$ aritmetička sredina svih izlaznih varijabli.

Stavi li se

(14) $$S^2 = \frac{1}{n-1} Q , \quad S_1^2 = \frac{1}{m-1} Q_1 , \quad S_2^2 = \frac{1}{n-m} Q_2 ,$$

može se reći da je S^2 korigirana varijanca svih mjerenja, S_1^2 je korigirana varijanca aritmetičkih sredina nizova, dok se S_2^2 može interpretirati kao pokazatelj prosječnog rasipanja unutar nizova.

Ako je hipoteza H_0 stvarno istinita, onda vrijedi:

(15) $$Y_{ij} \sim \mathrm{N}(\mu, \sigma^2) , \qquad i = 1, \ldots, m , \quad j = 1, \ldots, n_i$$

(16) $$\overline{Y}_i \sim \mathrm{N}(\mu, \frac{\sigma^2}{n_i}) , \qquad i = 1, \ldots, m ,$$

(17) $$\overline{Y} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) ,$$

pa na temelju onog što je navedeno u VI.4. (formula (53)), proizlazi

(18) $$\frac{1}{\sigma^2} Q = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1) ,$$

(19) $$\frac{1}{\sigma^2} Q_1 = \frac{m-1}{\sigma^2} S_1^2 \sim \chi^2(m-1) ,$$

(20) $$\frac{1}{\sigma^2} Q_2 = \frac{n-m}{\sigma^2} S_2^2 \sim \chi^2(n-m),$$

te činjenica da su Q_1 i Q_2 nezavisne slučajne varijable.

Primijeni li se, konačno, rezultat iz točke 8. u V.6, dobiva se

(21) $$V = \frac{S_1^2}{S_2^2} \sim \mathrm{F}(m-1, n-m),$$

pa će se vrijednost $v = \frac{s_1^2}{s_2^2}$, test-statistike V, uzeti kao kriterij za donošenje odluke o prihvaćanju ili odbacivanju hipoteze H_0. Da je ta vrijednost zaista prikladna za donošenje spomenute odluke, može se zaključiti iz činjenice (v. zad. 4) da je

(22) $$\mathrm{E}[S_1^2] = \sigma^2 + \frac{1}{m-1}\sum_{i=1}^{m} n_i \delta_i^2, \qquad \mathrm{E}[S_2^2] = \sigma^2,$$

bez obzira na hipotezu H_0. Ako je hipoteza H_0 istinita, onda je, dakako, $\mathrm{E}[S_1^2] = \sigma^2$, pa se može očekivati da će vrijednost v biti blizu jedinice, a ako hipoteza H_0 nije istinita, onda se može očekivati povećanje veličine s_1^2, pa stoga i omjera $v = \frac{s_1^2}{s_2^2}$.

Zato će se hipoteza H_0 odbaciti, ako se dobije prevelika vrijednost za v, tj. ako se dobije

$$v \geq \mathrm{F}_{m-1,n-m}^{-1}(1-\alpha),$$

gdje je $\alpha (0 < \alpha < 1)$ zadana razina značajnosti, a $F_{m-1,n-m}^{-1}$ inverzna funkcija od f.r.v. za F-razdiobu sa $(m-1, n-m)$ stupnjeva slobode (v. tabl. VII. u Dodatku).

Provedu li se odgovarajući računi na podacima iz 1. primjera, dobivaju se vrijednosti slučajnih varijabli iz (8)-(12), tj.

$$\overline{y}_1 = 7{,}0, \quad \overline{y}_2 = 7{,}5, \quad \overline{y}_3 = 7{,}6, \quad \overline{y} = 7{,}34,$$

$$q = 3{,}00, \quad q_1 = 1{,}16, \quad q_2 = q - q_1 = 1{,}84.$$

Na temelju (19) i (20) izračunava se

$$s_1^2 = 0{,}58, \qquad s_2^2 = 0{,}14,$$

iz čega proizlazi

$$v = 4{,}14.$$

Uzme li se razina značajnosti $\alpha = 0{,}05$, iz tabl. VII. u Dodatku odčitava se

$$F_{2,13}^{-1}(0{,}95) = 3{,}81,$$

pa se vidi da vrijednost (4,14) test-statistike V iz (21) pada u kritično područje $[3{,}81; \infty\rangle$, što znači da hipotezu o nepostojanju značajne razlike u potrošnji goriva promatranih maraka automobila ne treba prihvatiti.

Radi boljeg pregleda i veće jasnoće problema i njegovih rezultata, uobičajeno je da se proračun analize varijance prikazuje u obliku tzv. *ANOVA – tablice* (tablice analize varijance). U tabl. 2. prikazan je opći oblik ANOVA-tablice za jednofaktorski model analize varijance.

Tablica 2.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Korigirana varijanca	Vrijednost test-statistike
razlika među nizovima	$m-1$	$q_1 = \sum_{i=1}^{m} n_i(\overline{y}_i - \overline{y})^2$	$s_1^2 = \frac{1}{m-1} q_1$	$v = \frac{s_1^2}{s_2^2}$
slučajna greška	$n-m$	$q_2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$	$s_2^2 = \frac{1}{n-m} q_2$	
ukupno	$n-1$	$q = q_1 + q_2$		

3. Dvofaktorski aditivni model

Mnoge praktične situacije zahtijevaju da se promatra utjecaj dvaju faktora na ishod određene pojave. Tako se, na primjer, može postaviti teza da osim faktora marke automobila, na potrošnju goriva utječe i vozačko iskustvo. Stoga će se promatrati i faktor iskustvo, koji također može imati više razina, pa se prirodno nameće pitanje o postojanju ili nepostojanju značajnog utjecaja na potrošnju goriva jednog i drugog faktora, te o eventualnom postojanju međusobne interakcije između ta dva faktora.

2. primjer

Vozači su kategorizirani, prema vozačkom iskustvu, u pet razreda:

1. razred – početnici s vozačkim stažem manjim od 1 godine
2. razred – vozači sa stažem od 1 do 5 godina
3. razred – vozači sa stažem od 5 do 10 godina
4. razred – vozači sa stažem od 10 do 20 godina
5. razred – vozači sa stažem većim od 20 godina.

Slučajno se biraju po tri vozača iz svakog razreda. Jednom se daje automobil marke A, drugom marke B i trećem marke C, koje voze po istoj cesti i pritom se mjeri potrošnja goriva (litara na 100 kilometara). Dobiveni rezultati prikazani su u tabl. 3.

Tablica 3.

Marka automobila \ Iskustveni razred	1.	2.	3.	4.	5.	Aritmetička sredina
A	10,1	9,7	11,0	9,3	8,9	9,80
B	8,9	8,7	9,4	8,3	8,6	8,78
C	11,0	10,4	10,8	9,4	9,2	10,16
Aritmetička sredina	10,0	9,6	10,4	9,0	8,9	9,58

Smije li se, na temelju dobivenih podataka, zaključiti da I. faktor (marka automobila), sam za sebe, odnosno II. faktor (vozačko iskustvo), sam za sebe, ne utječe značajno na potrošnju goriva? Može se, također, postaviti i pitanje o eventualnom zajedničkom utjecaju (interakciji) obaju faktora na potrošnju goriva.

Da bi se dobio odgovor na postavljena pitanja, nužno je izgraditi odgovarajući matematički model, koji bi nas trebao uputiti i na organizaciju odgovarajućeg eksperimenta za prikupljanje potrebnih podataka.

Pretpostavimo da I. faktor općenito ima m_1 ($m_1 \geq 2$), a II. faktor m_2 ($m_2 \geq 2$) razina i da se raspolaže sa $n = m_1 m_2$ podataka y_{ij} ($i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$), gdje je y_{ij} vrijednost izlazne slučajne varijable Y_{ij} dobivena zbog djelovanja i-te razine I. faktora, j-te razine II. faktora i slučajne greške, tako da se može pisati

$$Y_{ij} = \mu_{ij} + \mathcal{E}_{ij} ,$$

pri čemu se pretpostavlja da su $\mathcal{E}_{ij}$ nezavisne slučajne varijable sa zajedničkom normalnom razdiobom $\mathrm{N}(0, \sigma^2)$. Može se, dakle, reći da je podatak y_{ij} dobiven mjerenjem slučajne varijable $Y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$.

Zapišimo mjerenja y_{ij} ($i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$) u obliku tablice 4.

Tablica 4.

Razine I. faktora \ Razine II. faktora	1	$\cdots$	j	$\cdots$	m_2
1	y_{11}	$\cdots$	y_{1j}	$\cdots$	y_{1m_2}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	y_{i1}	$\cdots$	y_{ij}	$\cdots$	y_{im_2}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
m_1	$y_{m_1 1}$	$\cdots$	$y_{m_1 j}$	$\cdots$	$y_{m_1 m_2}$

U 2. primjeru I. faktor je marka automobila, II. faktor vozačko iskustvo, $m_1 = 3$ i $m_2 = 5$, dok y_{ij} označuje potrošnju goriva automobila i-te marke, kada ga vozi vozač iz j-tog iskustvenog razreda.

Ustanovimo da je u svakom polju tabl. 4. zapisano jedno mjerenje slučajne varijable Y_{ij}. Može se, dakako, zamisliti i model u kojem se pretpostavlja više mjerenja slučajne varijable Y_{ij}, tj. realna situacija u kojoj je moguće izvesti više mjerenja izlazne varijable uz djelovanje i-te razine I. faktora i j-te razine II. faktora. Tada bi se u svakom polju tablice našlo i više od jednog podatka, što će se razmotriti kasnije (v. XIII.4).

Uvedimo oznake

(23) $$\mu = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \mu_{ij} ,$$

(24) $$\mu_i' = \frac{1}{m_2} \sum_{j=1}^{m_2} \mu_{ij} , \qquad \delta_i' = \mu_i' - \mu , \qquad i = 1, \ldots, m_1 ,$$

(25) $$\mu_j'' = \frac{1}{m_1} \sum_{i=1}^{m_1} \mu_{ij} , \qquad \delta_j'' = \mu_j'' - \mu , \qquad j = 1, \ldots, m_2 ,$$

pri čemu se μ može shvatiti kao opća srednja vrijednost, μ_i' kao srednja vrijednost izlazne varijable uz fiksiranu i-tu razinu I. faktora, a μ_j'' kao srednja vrijednost izlazne varijable uz fiksiranu j-tu razinu II. faktora. Zato se δ_i' zove ***glavni efekt*** i-te razine I. faktora, a δ_j'' glavni efekt j-te razine II. faktora.

Lako se provjerava (v. zad. 5) da vrijedi

(26) $$\sum_{i=1}^{m_1} \delta_i' = \sum_{j=1}^{m_2} \delta_j'' = 0 .$$

Uvede li se zapis

(27) $$\mu_{ij} = \mu + \delta_i' + \delta_j'' + \delta_{ij} ,$$

može se reći da je očekivana vrijednost u polju (i, j) tablice 4. rastavljena na zbroj u kojemu prvi član μ karakterizira opću srednju vrijednost izlazne varijable, član δ_i' glavni efekt i-te razine I. faktora, član δ_j'' glavni efekt j-te razine II. faktora, dok član δ_{ij} karakterizira doprinos međusobne interakcije i-te razine I. faktora i j-te razine II. faktora. Stoga se kao nul-hipoteze prirodno nameću hipoteze

(28) $$\mathrm{H}_{01} : \delta_1' = \cdots = \delta_{m_1}' = 0 ,$$

(29) $$\mathrm{H}_{02} : \delta_1'' = \cdots = \delta_{m_2}'' = 0 ,$$

(30) $$\mathrm{H}_{00} : \delta_{ij} = 0 , \quad \text{za} \quad i = 1, \ldots, m_1 , \quad j = 1, \ldots, m_2 ,$$

pri čemu se za svaku od njih, kao alternativna hipoteza, uzima da bar na jednom mjestu ne vrijedi znak jednakosti.

Testirati hipotezu H_{01} znači odgovoriti na pitanje je li utjecaj I. faktora na izlazne podatke značajan. Testiranjem hipoteze H_{02} dobiva se odgovor na pitanje da li je utjecaj II. faktora na izlazne podatke značajan, dok se testiranjem hipoteze H_{00} dobiva odgovor na pitanje da li postoji interakcija između I. i II. faktora, koja uzrokuje značajne promjene na izlaznim podacima. Budući da postojanje interakcije znatno otežava problem konstrukcije odgovarajućeg testa, najprije ćemo

razmotriti slučaj gdje se već u modelu pretpostavlja da interakcija ne postoji ($\delta_{ij} = 0$) i tada se govori o *aditivnom modelu*. U tom se slučaju modelska jednadžba zapisuje u obliku

$$Y_{ij} = \mu + \delta'_i + \delta''_j + \mathcal{E}_{ij}\,, \tag{31}$$

iz čega se razabire da I. i II. faktor imaju aditivni efekt na izlaznu varijablu.

U pronalaženju prikladnih test-statistika za testiranje hipoteza H_{01} i H_{02}, uz zadanu razinu značajnosti α, postupit će se slično kao u prethodnom poglavlju. Uvode se statistike

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2} Y_{ij} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \qquad n = m_1 m_2\,, \tag{32}$$

$$\overline{Y}'_i = \frac{1}{m_2}\sum_{j=1}^{m_2} Y_{ij} \sim \mathrm{N}\left(\mu + \delta'_i, \frac{\sigma^2}{m_2}\right), \qquad i = 1, \ldots, m_1\,, \tag{33}$$

$$\overline{Y}''_j = \frac{1}{m_1}\sum_{i=1}^{m_1} Y_{ij} \sim \mathrm{N}\left(\mu + \delta''_j, \frac{\sigma^2}{m_1}\right), \qquad j = 1, \ldots, m_2\,, \tag{34}$$

$$Q = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(Y_{ij} - \overline{Y})^2\,, \tag{35}$$

$$Q_1 = m_2\sum_{i=1}^{m_1}(\overline{Y}'_i - \overline{Y})^2\,, \tag{36}$$

$$Q_2 = m_1\sum_{j=1}^{m_2}(\overline{Y}''_j - \overline{Y})^2\,, \tag{37}$$

$$Q_3 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(Y_{ij} - \overline{Y}'_i - \overline{Y}''_j + \overline{Y})^2\,. \tag{38}$$

Vidi se da je

$\overline{Y}$ – aritmetička sredina svih mjerenja,
$\overline{Y}'_i$ – aritmetička sredina mjerenja dobivenih djelovanjem i-te razine I. faktora (i-tog retka tabl. 4),
$\overline{Y}''_j$ – aritmetička sredina mjerenja dobivenih djelovanjem j-te razine II. faktora (j-tog stupca tabl. 4),
Q – zbroj kvadrata odstupanja svih mjerenja od njihove aritmetičke sredine,
Q_1 – zbroj kvadrata odstupanja sredina redaka od zajedničke sredine,
Q_2 – zbroj kvadrata odstupanja sredina stupaca od zajedničke sredine,
Q_3 – rezidualni zbroj kvadrata.

Lako se pokazuje (v. zad. 7) da vrijedi

$$Q = Q_1 + Q_2 + Q_3\,, \tag{39}$$

što je analogon formule (22). Problem će se i ovaj puta rješavati razmatranjem odnosa između veličina Q_1, Q_2 i Q_3.

Iz pretpostavki modela te formula (35)-(38) proizlazi (v. zad. 8, 9. i 11)

$$\mathrm{E}[Q_1] = (m_1 - 1)\sigma^2 + m_2\sum_{i=1}^{m_1}(\delta'_i)^2\,, \tag{40}$$

$$\mathrm{E}[Q_2] = (m_2 - 1)\sigma^2 + m_1\sum_{j=1}^{m_2}(\delta''_j)^2\,, \tag{41}$$

$$\mathrm{E}[Q_3] = (m_1 - 1)(m_2 - 1)\sigma^2\,, \tag{42}$$

pa se vidi da je

$$S_3^2 = \frac{1}{(m_1 - 1)(m_2 - 1)}\, Q_3 \tag{43}$$

nepristrani procjenitelj za nepoznati parametar σ^2, bez obzira na hipoteze (28) i (29), dok će

$$S_1^2 = \frac{1}{m_1 - 1}\, Q_1 \tag{44}$$

biti nepristrani procjenitelj za σ^2 samo ako je hipoteza (28) stvarno istinita, a

$$S_2^2 = \frac{1}{m_2 - 1}\, Q_2 \tag{45}$$

bit će nepristrani procjenitelj za σ^2 samo ako je hipoteza (29) stvarno istinita. U protivnom se za S_1^2 i S_2^2 mogu očekivati veće vrijednosti, kako se razabire iz (40) i (41).

Zato se čini prikladnim uzeti, kao test-statistiku za testiranje hipoteze H_{01} da I. faktor nema značajni utjecaj na izlazne podatke, omjer

$$V_1 = \frac{S_1^2}{S_3^2}\,, \tag{46}$$

a kao test-statistiku za testiranje hipoteze H_{02} da II. faktor nema značajni utjecaj na izlazne podatke omjer

$$V_2 = \frac{S_2^2}{S_3^2}\,. \tag{47}$$

Prevelika vrijednost omjera $v_1 = \frac{s_1^2}{s_3^2}$, odnosno $v_2 = \frac{s_2^2}{s_3^2}$, upućivat će na odbacivanje hipoteze H_{01}, odnosno H_{02}.

Povoljna je okolnost da se može dokazati

$$V_1 \sim \mathrm{F}(r_1, s)\,, \qquad r_1 = m_1 - 1\,, \qquad s = (m_1 - 1)(m_2 - 1)\,, \tag{48}$$

odnosno

$$V_2 \sim \mathrm{F}(r_2, s)\,, \qquad r_2 = m_2 - 1\,, \tag{49}$$

što omogućuje da se odredi kritično područje, zadane razine značajnosti α, primjenom tablice (v. tabl. VII. u Dodatku) za F-razdiobu. Tako će se hipoteza H_{01} odbaciti ako se dobije

$$v_1 \geq F^{-1}_{r_1,s}(1-\alpha)\,, \tag{50}$$

dok će se hipoteza H_{02} odbaciti ako se dobije

$$v_2 \geq F^{-1}_{r_2,s}(1-\alpha)\,. \tag{51}$$

Sada možemo izvedene teorijske rezultate primijeniti na rješavanje praktičnog zadatka opisanog u 2. primjeru, gdje je $r_1 = 2$, $r_2 = 4$ i $s = 8$. Na temelju podataka iz tabl. 3. dobiva se

$$\overline{y} = 9{,}58\,, \quad q = 11{,}06\,, \quad q_1 = 5{,}12 \quad q_2 = 4{,}94\,,$$

iz čega se računa

$$q_3 = q - q_1 - q_2 = 1{,}00\,,$$

i dalje

$$s_1^2 = \frac{1}{m_1 - 1}\, q_1 = 2{,}56\,,$$
$$s_2^2 = \frac{1}{m_2 - 1}\, q_2 = 1{,}235\,,$$
$$s_3^2 = \frac{1}{(m_1 - 1)(m_2 - 1)}\, q_3 = 0{,}125\,,$$

iz čega se konačno dobiva $v_1 = 20{,}48$ i $v_2 = 9{,}88$.

Tablica 5.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Korigirana varijanca	Vrijednost test-statistike
razlike među recima (tabl. 4)	$m_1 - 1$	$q_1 = m_2 \sum_{i=1}^{m_1} (\overline{y}'_i - \overline{y})^2$	$s_1^2 = \frac{1}{m_1 - 1} q_1$	$v_1 = \frac{s_1^2}{s_3^2}$
razlike među stupcima (tabl. 4)	$m_2 - 1$	$q_2 = m_1 \sum_{j=1}^{m_2} (\overline{y}''_j - \overline{y})^2$	$s_2^2 = \frac{1}{m_2 - 1} q_2$	$v_2 = \frac{s_2^2}{s_3^2}$
slučajna greška	$(m_1 - 1)(m_2 - 1)$	$q_3 = q - q_1 - q_2$	$s_3^2 = \frac{1}{(m_1 - 1)(m_2 - 1)} q_3$	
ukupno	$m_1 m_2 - 1$	$q = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (y_{ij} - \overline{y})^2$		

Uzme li se $\alpha = 0{,}05$, iz tabl. VII. u Dodatku, odčitava se

$$F^{-1}_{2,8}(0{,}95) = 4{,}46\,, \qquad F^{-1}_{4,8}(0{,}95) = 3{,}84\,,$$

pa se vidi da $v_1 = 20{,}48$ pada u kritično područje $[4{,}46; \infty\rangle$, a isto tako i $v_2 = 9{,}88$ pada u kritično područje $[3{,}84; \infty\rangle$, što znači da obje hipoteze H_{01} i H_{02} treba odbaciti. Podaci iz tabl. 3. upućuju, dakle, na zaključak da i marka automobila i vozačko iskustvo utječu na potrošnju goriva.

Slično kao u jednofaktorskom modelu analize varijance, i u dvofaktorskome aditivnom modelu proračun relevantnih veličina obično se prikazuje u obliku ANOVA-tablice dvofaktorskoga aditivnog modela (v. tabl. 5).

4. Opći dvofaktorski model

Da bi se mogla testirati i hipoteza H_{00}, izražena u (30), koja izražava odsustvo interaktivnog utjecaja obaju faktora, nužno je imati više od jednog podatka u svakom polju tabl. 4. Kako to izgleda u praksi ilustrirat ćemo opet na primjeru potrošnje goriva (1. i 2. primjer).

3. primjer

Usvajamo sve pretpostavke iz 2. primjera, osim što umjesto po tri vozača iz svakog razreda slučajno biramo devet vozača i trojica voze automobil marke A, trojica marke B i trojica marke C, pri čemu se mjeri odgovarajuća potrošnja goriva (litara na 100 km). Dobiveni rezultati prikazani su u tabl. 6.

Tablica 6.

Marka automobila \ Iskustveni razred	1.	2.	3.	4.	5.
A	10,0 10,4 10,2	9,8 9,8 9,5	11,1 11,3 11,2	9,1 9,4 9,4	9,0 8,6 9,1
B	9,0 9,4 8,9	8,5 8,8 8,8	9,6 9,9 9,9	8,9 9,0 8,8	9,4 9,4 9,1
C	10,9 10,8 10,4	10,2 10,0 10,4	10,7 10,9 10,5	9,0 8,6 8,5	10,1 10,2 10,3

Odmah vidimo da se u svakom polju tabl. 6. nalaze tri ($l = 3$) vrijednosti i svaku od njih smatramo rezultatom mjerenja slučajne varijable $Y_{ijk} = \mu_{ij} + \mathcal{E}_{ijk}$, gdje je μ_{ij} neslučajna veličina koja karakterizira djelovanje i-te razine I. faktora (marka automobila) i j-te razine II. faktora (vozačko iskustvo), dok je $\mathcal{E}_{ijk}$ slučajna varijabla koja karakterizira slučajnu grešku k-tog mjerenja u polju (i, j).

Matematički opis dvofaktorskog modela analize varijance s više podataka po polju općenito je izražen jednadžbom

$$(52)\qquad Y_{ijk} = \mu_{ij} + \mathcal{E}_{ijk}\,, \qquad i = 1,\ldots,m_1\,,\quad j = 1,\ldots,m_2\,,\quad k = 1,\ldots,l\,,$$

gdje su $\mathcal{E}_{ijk}$ nezavisne slučajne varijable sa zajedničkom normalnom razdiobom $N(0,\sigma^2)$. Izlaznu vrijednost y_{ijk} tretiramo kao jednu od l vrijednosti u polju (i,j) tablice podataka. Smatra se, dakle, da je y_{ijk} rezultat djelovanja i-te razine I. faktora, j-te razine II. faktora, međusobne interakcije obaju faktora i slučajne greške normalne razdiobe s očekivanjem nula i varijancom σ^2. To se može izraziti, kao i u (27), tako da se stavi

$$(53)\qquad \mu_{ij} = \mu + \delta_i' + \delta_j'' + \delta_{ij}\,,$$

gdje je, slično kao u (23) – (26),

$$(54)\qquad \mu = \frac{1}{m_1 m_2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mu_{ij}\,,$$

$$(55)\qquad \mu_i' = \frac{1}{m_2}\sum_{j=1}^{m_2}\mu_{ij}\,,\qquad \delta_i' = \mu_i' - \mu\,,\quad i = 1,\ldots,m_1\,,$$

$$(56)\qquad \mu_j'' = \frac{1}{m_1}\sum_{i=1}^{m_1}\mu_{ij}\,,\qquad \delta_j'' = \mu_j'' - \mu\,,\quad j = 1,\ldots,m_2\,,$$

$$(57)\qquad \sum_{i=1}^{m_1}\delta_i' = \sum_{j=1}^{m_2}\delta_j'' = \sum_{i=1}^{m_1}\delta_{ij} = \sum_{j=1}^{m_2}\delta_{ij} = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\delta_{ij} = 0\,.$$

Interpretacija veličina μ, μ_i', μ_j'', δ_i', δ_j'' i δ_{ij} ista je kao u XIII.3. Sada je još očiglednije zašto se δ_i' zove glavni efekt i-te razine I. faktora, a δ_j'' glavni efekt j-te razine II. faktora, dok se δ_{ij} zove *interakcijski efekt*, čiji utjecaj se također može proučavati u ovom modelu analize varijance.

Glavni je problem i u ovom modelu da se definiraju prikladne test-statistike za testiranje hipoteza (28), (29) i (30). Ideja vodilja i ovaj je puta rastavljanje ukupnog rasipanja podataka na komponente, što je tipično za sve modele analize varijance. U tu svrhu uvode se statistike

$$(58)\qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{l} Y_{ijk} \sim N\left(\mu, \frac{\sigma^2}{n}\right)\,,\qquad n = m_1 m_2 l\,,$$

$$(59)\qquad \overline{Y}_i' = \frac{1}{m_2 l}\sum_{j=1}^{m_2}\sum_{k=1}^{l} Y_{ijk} \sim N\left(\mu + \delta_i', \frac{\sigma^2}{m_2 l}\right)\,,\qquad i = 1,\ldots,m_1\,,$$

$$(60)\qquad \overline{Y}_j'' = \frac{1}{m_1 l}\sum_{i=1}^{m_1}\sum_{k=1}^{l} Y_{ijk} \sim N\left(\mu + \delta_j'', \frac{\sigma^2}{m_1 l}\right)\,,\qquad j = 1,\ldots,m_2\,,$$

$$(61)\qquad \overline{Y}_{ij} = \frac{1}{l}\sum_{k=1}^{l} Y_{ijk} \sim N\left(\mu + \delta_i' + \delta_j'' + \delta_{ij}, \frac{\sigma^2}{l}\right)\,,\qquad \begin{array}{l} i = 1,\ldots,m_1\,,\\ j = 1,\ldots,m_2\,,\end{array}$$

$$(62)\qquad Q = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{l}(Y_{ijk} - \overline{Y})^2\,,$$

$$(63)\qquad Q_1 = m_2 l\sum_{i=1}^{m_1}(\overline{Y}_i' - \overline{Y})^2\,,$$

$$(64)\qquad Q_2 = m_1 l\sum_{j=1}^{m_2}(\overline{Y}_j'' - \overline{Y})^2\,,$$

$$(65)\qquad Q_{12} = l\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(\overline{Y}_{ij} - \overline{Y}_i' - \overline{Y}_j'' + \overline{Y})^2\,,$$

$$(66)\qquad Q_3 = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{l}(Y_{ijk} - \overline{Y}_{ij})^2\,.$$

Definirane veličine možemo interpretirati ovako:

$\overline{Y}$ – aritmetička sredina svih mjerenja,
$\overline{Y}_i'$ – aritmetička sredina svih mjerenja iz i-tog retka,
$\overline{Y}_j''$ – aritmetička sredina svih mjerenja iz j-tog stupca,
$\overline{Y}_{ij}$ – aritmetička sredina svih mjerenja iz (i,j)-tog polja,
Q – zbroj kvadrata odstupanja svih mjerenja od njihove aritmetičke sredine,
Q_1 – zbroj kvadrata odstupanja sredina redaka od zajedničke sredine,
Q_2 – zbroj kvadrata odstupanja sredina stupaca od zajedničke sredine,
Q_{12} – interakcijski zbroj kvadrata,
Q_3 – zbroj kvadrata odstupanja mjerenja od odgovarajućih sredina u polju.

Pokazuje se (v. zad. 14) da vrijedi

$$(67)\qquad Q = Q_1 + Q_2 + Q_{12} + Q_3\,,$$

što je analogon jednakosti (39) u dvofaktorskom aditivnom modelu, uz primjedbu da se u ovom slučaju pojavljuje i član Q_{12} koji karakterizira interakciju I. i II. faktora.

Posebno je važno da se može dokazati (v. zad. 16,17,18. i 19) da vrijedi

$$(68)\qquad E[Q_1] = (m_1 - 1)\sigma^2 + m_2 l\sum_{i=1}^{m_1}(\delta_i')^2\,,$$

$$(69)\qquad E[Q_2] = (m_2 - 1)\sigma^2 + m_1 l\sum_{j=1}^{m_2}(\delta_j'')^2\,,$$

$$(70)\qquad E[Q_{12}] = (m_1 - 1)(m_2 - 1)\sigma^2 + 2l\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\delta_{ij}(\delta_i' + \delta_j'' + \delta_{ij})\,,$$

(71) $$\mathrm{E}[Q_3] = m_1 m_2 (l-1)\sigma^2 ,$$

pa se vidi da je

(72) $$S_3^2 = \frac{1}{m_1 m_2 (l-1)} Q_3$$

nepristrani procjenitelj za nepoznati parametar σ^2, bez obzira na hipoteze (28), (29) i (30), dok će

(73) $$S_1^2 = \frac{1}{m_1 - 1} Q_1 ,$$

kako se razabire iz (68), biti nepristrani procjenitelj za σ^2 samo ako je hipoteza (28) (I. faktor ne utječe na izlazne podatke) stvarno istinita. Isto tako će

(74) $$S_2^2 = \frac{1}{m_2 - 1} Q_2$$

biti nepristrani procjenitelj za σ^2 samo ako je hipoteza (29) (II. faktor ne utječe na izlazne podatke) stvarno istinita, što se razabire iz (69). Iz (70) se, pak, razabire da će

(75) $$S_{12}^2 = \frac{1}{(m_1 - 1)(m_2 - 1)} Q_{12}$$

biti nepristrani procjenitelj za σ^2 samo ako je hipoteza (30) (ne postoji interakcija obaju faktora) stvarno istinita.

Ako su hipoteze (28), (29) i (30) doista neistinite, onda se mogu očekivati, kako se vidi iz (68), (69) i (70), veće vrijednosti statistika S_1^2, S_2^2 i S_{12}^2 nego u slučaju stvarne istinitosti navedenih hipoteza. Sve to sugerira da se

(76) $$V_1 = \frac{S_1^2}{S_3^2} \sim \mathrm{F}(m_1 - 1 , m_1 m_2 (l-1))$$

uzme kao test-statistika za testiranje hipoteze H_{01}, da se

(77) $$V_2 = \frac{S_2^2}{S_3^2} \sim \mathrm{F}(m_2 - 1 , m_1 m_2 (l-1))$$

uzme kao test-statistika za testiranje hipoteze H_{02}, te da se test-statistika

(78) $$V_{12} = \frac{S_{12}^2}{S_3^2} \sim \mathrm{F}((m_1 - 1)(m_2 - 1) , m_1 m_2 (l-1))$$

uzme za testiranje hipoteze H_{12}.

Dobije li se vrijednost $v_1 = \frac{s_1^2}{s_3^2}$, odnosno $v_2 = \frac{s_2^2}{s_3^2}$, odnosno $v_{12} = \frac{s_{12}^2}{s_3^2}$, mnogo veća od jedinice, to će nas uputiti na odbacivanje hipoteze H_{01}, odnosno H_{02}, odnosno H_{12}. Uzme li se razina značajnosti α i dobije

(79) $$v_1 \geq F_{r_1,s}^{-1}(1-\alpha) , \qquad r_1 = m_1 - 1 , \qquad s = m_1 m_2 (l-1) ,$$

hipoteza H_{01} će se odbaciti. Dobije li se na temelju danih podataka

(80) $$v_2 \geq F_{r_2,s}^{-1}(1-\alpha) , \qquad r_2 = m_2 - 1 ,$$

hipoteza H_{02} će se dobaciti, a dobije li se

(81) $$v_{12} \geq F_{r_{12},s}^{-1}(1-\alpha) , \qquad r_{12} = (m_1 - 1)(m_2 - 1) ,$$

odbacit će se hipoteza H_{12}.

Proračun veličina v_1, v_2 i v_{12} obično se prikazuje u obliku ANOVA-tablice dvofaktorskog modela s više podataka po polju (v. tabl. 7).

Tablica 7.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja
razlike među recima	$m_1 - 1$	$q_1 = m_2 l \sum_{i=1}^{m_1} (\overline{y}_i' - \overline{y})^2$
razlike među stupcima	$m_2 - 1$	$q_2 = m_1 l \sum_{j=1}^{m_2} (\overline{y}_j'' - \overline{y})^2$
interakcija	$(m_1 - 1)(m_2 - 1)$	$q_{12} = l \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (\overline{y}_{ij} - \overline{y}_i' - \overline{y}_j'' + \overline{y})^2$
slučajna greška	$m_1 m_2 (l-1)$	$q_3 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{l} (y_{ijk} - \overline{y}_{ij})^2$
ukupno	$m_1 m_2 l - 1$	$q = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \sum_{k=1}^{l} (y_{ijk} - \overline{y})^2$

Izvor rasipanja	Korigirana varijanca	Vrijednost test- -statistike
razlike među recima	$s_1^2 = \frac{1}{m_1 - 1} q_1$	$v_1 = \frac{s_1^2}{s_3^2}$
razlike među stupcima	$s_2^2 = \frac{1}{m_2 - 1} q_2$	$v_2 = \frac{s_2^2}{s_3^2}$
interakcija	$s_{12}^2 = \frac{1}{(m_1 - 1)(m_2 - 1)} q_{12}$	$v_{12} = \frac{s_{12}^2}{s_3^2}$
slučajna greška	$s_3^2 = \frac{1}{m_1 m_2 (l-1)} q_3$	

Na temelju podataka iz 3. primjera (tabl. 6) dobiva se pripadna ANOVA--tablica (tabl. 8).

Tablica 8.

Izvor rasipanja	Broj stupnjeva slobode	Zbroj kvadrata odstupanja	Korigirana varijanca	Vrijednost test--statistike
razlike među recima	2	7,156	3,578	92,534
razlike među stupcima	4	13,148	3,287	85,009
interakcija	8	6,604	0,8255	21,349
slučajna greška	30	1,160	0,0387	
ukupno	44	28,068		

5. Testiranje hipoteza o adekvatnosti modela

Globalno gledajući na problematiku regresijske analize i analize varijance može se ustanoviti da su razvijeni različiti matematički modeli, u kojima postoje određene pretpostavke (tip regresijske funkcije, nezavisnost i normalnost slučajnih grešaka, odsustvo interakcije i sl.), koje se unaprijed usvajaju i na koje redovito ne utječu dani empirijski podaci. Mogu se, dapače, na iste podatke primijeniti različiti modeli.

Imajući, na primjer, na umu različite regresijske modele (linearni, polinomski, eksponencijalni i sl.), očigledno se može pokušati istim podacima prilagoditi, recimo, linearni i polinomski regresijski model. Na temelju same teorije modela ne može se egzaktno zaključiti o većoj ili manjoj prikladnosti (adekvatnosti) izabranog modela za opisivanje promatranoga praktičnog fenomena. Stoga se nameće zadatak nalaženja određenog postupka za ocjenu adekvatnosti usvojenog modela.

U svim razmatranim modelima regresije i analize varijance temeljna je pretpostavka da slučajna greška ima normalnu razdiobu $N(0,\sigma^2)$. U nešto pojednostavnjenom obliku opća modelska jednadžba glasi

(82) $$Y_i = \mu_i + \mathcal{E}_i\,, \qquad i = 1,\ldots,n\,,$$

pri čemu se pretpostavlja da su $\mathcal{E}_1,\ldots,\mathcal{E}_n$ nezavisne slučajne varijable, a $\mu_i (i = 1,\ldots,n)$ i σ^2 nepoznati parametri. U tim modelima razrađene su metode (redovito je to MNK) za dobivanje odgovarajućih procjenitelja za μ_i i σ^2. Ako je broj podataka n dovoljno velik, onda se može uzeti da je nepoznati parametar μ_i približno jednak vrijednosti $\widehat{\mu}_i$ pripadnog procjenitelja, tj. može se pisati $\mu_i \approx \widehat{\mu}_i$, a također i $\sigma^2 \approx \widehat{\sigma}^2$, gdje je $\widehat{\sigma}^2$ odgovarajuća procjena za σ^2.

U regresijskim je modelima $\widehat{\mu}_i = \mu_{\widehat{a}}(x_i)$ ($\widehat{a}$ je MNK-procjena za nepoznati parametar a regresijske funkcije), dok je u ANOVA-modelima $\widehat{\mu}_i$ aritmetička sredina onih vrijednosti izlazne varijable na koje djeluje određena kombinacija razina promatranih faktora.

U svim modelima se nepristrana procjena za nepoznati parametar σ^2 označuje sa s^2, pa se za velike n može pisati $\sigma^2 \approx s^2$.

U uvjetima adekvatnosti izabranog modela i velikog broja podataka niz

(83) $$z_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\sigma}}\,, \qquad i = 1,\ldots,n\,,$$

trebao bi se ponašati kao niz nezavisnih mjerenja slučajne varijable $Z \sim N(0,1)$, tj. kao vrijednost slučajnog uzorka iz standardne normalne razdiobe.

Prema tome, testiranje hipoteze o adekvatnosti modela (82), tj. o valjanosti pretpostavke da se slučajna greška podvrgava normalnoj razdiobi $N(0,\sigma^2)$, svodi se na zadatak o testiranju hipoteze H_0 da podaci (83) potječu iz standardne normalne razdiobe, što se može riješiti hikvadrat-testom (v. IX.3) ili KS-testom (v. X.2).

Ako su dani podaci takvi da testiranje završi odbacivanjem hipoteze H_0, onda to upućuje na sumnju u valjanost pretpostavke da greška ima normalnu razdiobu s očekivanjem nula i konstantnom varijancom σ^2. Može se, naime, desiti da varijanca greške nije konstantna veličina (*heteroscedasticity*), već ovisi o vrijednostima ulazne varijable i tada, naravno dolazi u pitanje adekvatnost modela, pa treba potražiti adekvatniji model (v. [33]).

Za testiranje hipoteze o prikladnosti izabranog tipa regresijske funkcije (afina, polinomska, eksponencijalna i sl.), može poslužiti postupak opisan u jednofaktorskom modelu analize varijance.

Pretpostavimo, stoga, da se u jednodimenzionalnome regresijskom modelu sa m ($m > 2$) različitih podataka o ulaznoj varijabli, za svaki x_i ($i = 1,\ldots,m$) raspolaže sa n_i odgovarajućih podataka y_{ij} ($j = 1,\ldots,n_i$) o izlaznoj slučajnoj varijabli. Podaci su shematski prikazani u tabl. 9.

Tablica 9.

i	Vrijednost ulazne varijable	Vrijednost izlazne varijable
1	x_1	$y_{11},\ldots,y_{1j},\ldots,y_{1n_1}$
$\vdots$	$\vdots$	$\vdots$
i	x_i	$y_{i1},\ldots,y_{ij},\ldots,y_{in_i}$
$\vdots$	$\vdots$	$\vdots$
m	x_m	$y_{m1},\ldots,y_{mj},\ldots,y_{mn_m}$

Matematički model za opisanu situaciju može se izraziti jednadžbom

(84) $$Y_{ij} = \mu_a(x_i) + \mathcal{E}_{ij}\,, \qquad i = 1,\ldots,m\,, \quad j = 1,\ldots,n_i\,,$$

što znači da se podatak y_{ij} može tretirati kao rezultat mjerenja slučajne varijable $Y_{ij} \sim \mathrm{N}(\mu_{\boldsymbol{a}}(x_i), \sigma^2)$.

Odmah se uočava velika sličnost s modelskom jednadžbom (2), samo što, umjesto μ_i u (2), u (84) stoji vrijednost regresijske funkcije $\mu_{\boldsymbol{a}}(x_i)$.

Neka je $\widehat{\boldsymbol{a}}$ MNK-procjena za nepoznati parametar $\boldsymbol{a}$ i neka je $\overline{y}_i = \dfrac{1}{n_i}\sum\limits_{j=1}^{n_i} y_{ij}$, pa na temelju relacije (15) iz II.3. slijedi da, za svaki $i = 1, \ldots, m$ vrijedi

$$\sum_{j=1}^{n_i}[y_{ij} - \mu_{\widehat{\boldsymbol{a}}}(x_i)]^2 \geq \sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 , \tag{85}$$

pri čemu znak jednakosti vrijedi onda i samo onda ako je $\mu_{\widehat{\boldsymbol{a}}}(x_i) = \overline{y}_i$, tj. ako se vrijednost procjene regresijske funkcije u točki x_i podudara s aritmetičkom sredinom podataka o izlaznoj varijabli za vrijednost x_i ulazne varijable.

Zbrajanjem nejednakosti (85) po $i = 1, \ldots, m$, dobiva se

$$\widehat{q} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}[y_{ij} - \mu_{\widehat{\boldsymbol{a}}}(x_i)]^2 \geq \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 = q_2 . \tag{86}$$

Kada bismo u (86) imali jednakost, mogli bismo smatrati da je izabrana prikladna regresijska funkcija, jer je zbroj kvadrata odstupanja $\widehat{q}$ najmanji mogući, pa se može smatrati i da je izabrani regresijski model adekvatan empirijskim podacima. Dobije li se $\widehat{q}$ mnogo veće od q_2, treba posumnjati u ispravnost izbora regresijskog modela.

Za egzaktno rješenje problema treba uočiti da se $\widehat{q}$ može tretirati kao vrijednost statistike $\widehat{Q} = (n - r)S^2$, gdje je S^2 definirano u XII.3. (relacija (34)), dok se q_2 može tretirati kao vrijednost statistike Q_2, definirane relacijom (12). U XII.6. (fundamentalni teorem) pokazano je da $\dfrac{1}{\sigma^2}\widehat{Q} = \dfrac{n-r}{\sigma^2}S^2 \sim \chi^2(n-r)$, dok je relacijom (20) iskazano da $\dfrac{1}{\sigma^2}Q_2 = \dfrac{n-m}{\sigma^2}S_2^2 \sim \chi^2(n-m)$, pa se, na temelju točke 8. iz V.6, može zaključiti da

$$V = \frac{n-m}{n-r}\,\frac{\widehat{Q}}{Q_2} = \frac{S^2}{S_2^2} \sim \mathrm{F}(n-r, n-m) . \tag{87}$$

Sada se vidi da se slučajna varijabla V može uzeti kao test-statistika za testiranje hipoteze H_0 : izabrani regresijski model je adekvatan, prema alternativnoj hipotezi da nije adekvatan. Dobije li se prevelika vrijednost omjera

$$v = \frac{s^2}{s_2^2} = \frac{n-m}{n-r}\,\frac{\widehat{q}}{q_2} , \tag{88}$$

hipotezu H_0 treba odbaciti.

Uzme li se razina značajnosti α, hipoteza H_0 će se odbaciti kada se dobije

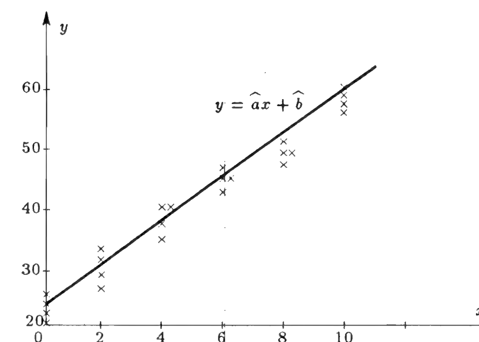
$$v \geq F^{-1}_{n-r,n-m}(1-\alpha) . \tag{89}$$

4. primjer

Mjerenjem određenih fizikalnih veličina dobivena je ova tablica podataka:

Tablica 10.

i	x_i	y_{ij}			
1	0	20	22	24	26
2	2	28	30	32	34
3	4	36	38	40	40
4	6	43	45	45	46
5	8	48	50	50	52
6	10	53	54	55	56



Slika 38. Grafički prikaz podataka iz tabl. 10.

Pogledamo li sl. 38. odmah se nameće ideja da se uzme regresijska funkcija oblika $\mu(x) = ax + b$. Primijene li se formule (4), (5) i (16) iz XI. poglavlja, dobivaju se vrijednosti procjena nepoznatih parametara a, b i σ^2, tj.

$$\widehat{a} = 3{,}15 , \qquad \widehat{b} = 24{,}5 , \qquad s^2 = 4{,}65 . \tag{90}$$

Uspoređujući tabl. 9. i 10, vidi se da je u promatranom primjeru $n = 24$, $m = 6$, $r = 2$ i $n_i = 4$ $(i = 1, \ldots, 6)$, pa se primjenom (86), (90) i podataka iz tabl. 10, dobiva

$$\widehat{q} = 102{,}35 , \qquad q_2 = 68{,}75 ,$$

dok se primjenom (88) dobiva

$$v = 1{,}22 .$$

Uzme li se razina značajnosti $\alpha = 0{,}05$, iz tabl. VII. u Dodatku očitava se

$$F^{-1}_{22,18}(0{,}95) = 2{,}17\,,$$

pa se vidi da vrijednost test-statistike (1,22) ne pada u kritično područje $[2{,}17;\infty\rangle$, što upućuje na prihvaćanje hipoteze H_0 da je izabrani regresijski model (afina regresijska funkcija) adekvatan.

Odmah primijetimo da time nisu isključeni svi drugi regresijski modeli, jer se, dakako, može dogoditi da i za neki drugi regresijski model opisani test omogući prihvaćanje hipoteze H_0. Smisao je ovoga testa, zapravo, da se isključe neadekvatni modeli.

6. Durbin-Watsonov test

Već je istaknuto da je u općoj modelskoj jednadžbi (82) ključna pretpostavka da su $\mathcal{E}_1,\ldots,\mathcal{E}_n$ nezavisne slučajne varijable. Ako ta pretpostavka nije ispunjena, narušavaju se bitni zaključci izvedeni u razmotrenim modelima regresijske analize. Tako MNK-procjenitelji nepoznatih parametara regresijske funkcije postaju slabo efikasni, a postupci za dobivanje intervala povjerenja i za testiranja hipoteza glede parametara modela ostaju bez teorijskog utemeljenja, pa njihova primjena može dovesti do vrlo pogrešnih zaključaka.

Stoga se prirodno nameće ideja da se konstruira, ako je moguće, test za provjeru nezavisnosti grešaka u danom nizu mjerenja. Iz (82) se vidi da je $\mathcal{E}_i = Y_i - \mu_i$, pa je razumljivo da se niz slučajnih varijabli $\mathcal{E}_i$ $(i = 1,\ldots,n)$ proučava pomoću niza reziduuma

$$\widehat{\varepsilon}_i = y_i - \widehat{\mu}_i\,, \qquad i = 1,\ldots,n\,, \tag{91}$$

uvedenih u XI.3. i XII.4. Veličina $\widehat{\varepsilon}_i$ pokazuje razliku između izmjerene vrijednosti y_i izlazne slučajne varijable Y_i i procjene $\widehat{\mu}_i$ nepoznate modelske vrijednosti μ_i pri i-tom mjerenju. Postavlja se, dakle, zadatak da se na temelju niza podataka (91) otkrije zavisnost, odnosno utvrdi nezavisnost niza slučajnih grešaka $\mathcal{E}_1,\ldots,\mathcal{E}_n$.

Da bi se to postiglo nužno je usvojiti neke dodatne pretpostavke. Poznata je činjenica (v. V.5) da se za niz slučajnih varijabli normalne razdiobe pojam nezavisnosti podudara s pojmom nekoreliranosti, pa se u tom slučaju zavisnost slučajnih varijabli može proučavati pomoću kovarijance, odnosno pomoću koeficijenata korelacije. U tu svrhu prikladno je niz slučajnih grešaka $\mathcal{E}_i$ $(i = 1,\ldots,n)$ interpretirati kao vremenski niz slučajnih varijabli, gdje se pretpostavlja da koreliranost opada s vremenskom razdaljinom, tj. da vrijedi

$$\mathrm{Cov}(\mathcal{E}_i,\mathcal{E}_{i-j}) = \varrho^j\sigma^2\,, \qquad j = 1,2,\ldots\,, \tag{92}$$

odnosno da za odgovarajuće koeficijente korelacije vrijedi

$$r_j = \varrho^j\,, \qquad j = 1,2,\ldots\,, \tag{93}$$

pri čemu je $\varrho (0 \le |\varrho| \le 1)$ zadani broj koji pokazuje stupanj koreliranosti greške u trenutku i i greške u pretodnom trenutku $i-1$. Pretpostavlja se, zapravo, da je koeficijent korelacije između greške i-tog i $(i-j)$-tog mjerenja ϱ^j, tj. da ne ovisi o trenutku i, već samo o vremenskoj udaljenosti j između dva uočena mjerenja u danom nizu mjerenja.

Ako je, na primjer, koeficijent korelacije između grešaka susjednih mjerenja $\varrho = 0{,}2$, onda će za mjerenja udaljena $j = 2$ koraka u nizu iznositi $\varrho^2 = 0{,}04$, za $j = 3$ koraka iznositi će $\varrho^3 = 0{,}008$ itd.

Ostalo je još da se definira prikladna test-statistika, koja će reagirati na odstupanje od nul-hipoteze $\mathrm{H}_0 : \varrho = 0$. Pokazalo se da je to *Durbin-Watsonova statistika* (DW-statistika, v. [10])

$$\widehat{D} = \frac{\sum\limits_{i=2}^{n}(\widehat{\mathcal{E}}_i - \widehat{\mathcal{E}}_{i-1})^2}{\sum\limits_{i=1}^{n}\widehat{\mathcal{E}}_i^2}\,, \tag{94}$$

gdje je $\widehat{\mathcal{E}}_i$ slučajna varijabla s vrijednostima $\widehat{\varepsilon}_i$ definiranim u (91). Za daljnja razmatranja bitna je pretpostavka da je procjena $\widehat{\mu}_i$ dobivena metodom najmanjih kvadrata.

Da je DW-statistika $\widehat{D}$ zaista prikladna za testiranje nul-hipoteze $\mathrm{H}_0 : \varrho = 0$, prema alternativnoj hipotezi $H_1 : \varrho \neq 0$ (ili $\varrho < 0$, ili $\varrho > 0$) može se vidjeti tako da se razvijeno napiše izraz za vrijednost $\widehat{d}$ test-statistike $\widehat{D}$. Iz (94) se vidi da je

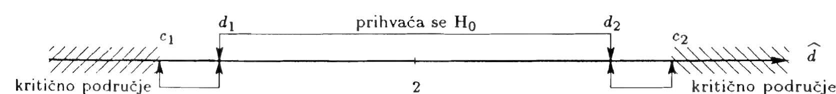
$$\widehat{d} = \frac{\sum\limits_{i=2}^{n}\widehat{\varepsilon}_i^2}{\sum\limits_{i=1}^{n}\widehat{\varepsilon}_i^2} + \frac{\sum\limits_{i=2}^{n}\widehat{\varepsilon}_{i-1}^2}{\sum\limits_{i=1}^{n}\widehat{\varepsilon}_i^2} - 2\,\frac{\sum\limits_{i=2}^{n}\widehat{\varepsilon}_i\widehat{\varepsilon}_{i-1}}{\sum\limits_{i=1}^{n}\widehat{\varepsilon}_i^2}\,. \tag{95}$$

Ako je n veliko, onda će prvi i drugi član desne strane u (95) približno iznositi 1, dok je treći član približno jednak $2\widehat{\varrho}$, gdje je $\widehat{\varrho}$ vrijednost uzoračkog koeficijenta korelacije grešaka susjednih mjerenja. Stoga se, umjesto (95), može pisati da je

$$\widehat{d} \approx 2(1-\widehat{\varrho})\,. \tag{96}$$

Kada su susjedne greške mjerenja nekorelirane, tj. $\varrho = 0$, tada se i za $\widehat{\varrho}$ očekuje vrijednost bliska nuli, osnosno za $\widehat{d}$ se očekuje vrijednost bliska 2. Dobije li se kao vrijednost $\widehat{d}$ test-statistike $\widehat{D}$ broj mnogo veći od 2, hipoteza da su greške $\mathcal{E}_1,\ldots,\mathcal{E}_n$ nekorelirane će se odbaciti i prihvatiti alternativna hipoteza da postoji značajna pozitivna korelacija u danom nizu grešaka. Dobije li se pak broj mnogo manji od 2, usvojit će se hipoteza da postoji značajna negativna korelacija u danom nizu grešaka.

Za egzaktno utvrđivanje granica kritičnog područja trebalo bi poznavati razdiobu vjerojatnosti test-statistike $\widehat{D}$ iz (94). Tu se pojavljuju velike poteškoće, jer i u uvjetima istinitosti nul-hipoteze ta razdioba vjerojatnosti ovisi o vrijednostima ulaznih (neslučajnih) varijabli (faktora), što bi u praktičnim primjenama zahtijevalo odgovarajuće tablice za svaku moguću vrijednost ulaznih varijabli. Olakšavajuća je okolnost da razdioba vjerojatnosti test-statistike $\widehat{D}$ nije previše osjetljiva na promjene ulaznih varijabli, tako da se područje vrijednosti $\widehat{d}$ može podijeliti na tri dijela, kako je skicirano na sl. 39.



Slika 39. Skica odlučivanja u DW-testu

Shema odlučivanja u Durbin-Watsonovu testu (DW-testu) prikazana je tablicom 11.

Tablica 11.

Vrijednost test-statistike	Odluka
$\widehat{d} \in \langle d_1, d_2 \rangle$	prihvaća se H_0 : $\varrho = 0$
$\widehat{d} \in \langle -\infty, c_1 \rangle$	odbacuje se H_0: $\varrho = 0$ i prihvaća H_1: $\varrho < 0$
$\widehat{d} \in \langle c_2, \infty \rangle$	odbacuje se H_0: $\varrho = 0$ i prihvaća H_1: $\varrho > 0$
$\widehat{d} \in [c_1, d_1] \cup [d_2, c_2]$	ne donosi se odluka

Budući da razdioba vjerojatnosti test-statistike $\widehat{D}$ bitno ovisi o broju r ulaznih varijabli modela, izrađene su tablice (v. tabl. XII. u Dodatku), koje omogućuju određivanje brojeva c_1 i d_1 za dani r i danu razinu značajnosti α, dok je

$$c_2 = 4 - c_1 \,, \qquad d_2 = 4 - d_1 \,. \tag{97}$$

5. primjer

Uzmemo li podatke iz 4. primjera, u kojem smo imali regresijsku funkciju

$$\widehat{\mu}(x) = 3{,}15\, x + 24{,}5 \,,$$

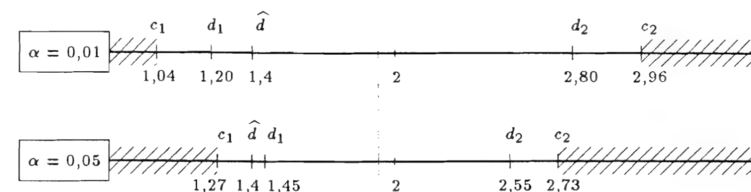
možemo načiniti tablicu (tabl. 12) u kojoj su prikazane vrijednosti regresijske funkcije $\widehat{\mu}_i(x_i) = \widehat{\mu}_i$, reziduuma $\widehat{\varepsilon}_i = y_i - \widehat{\mu}_i$ $(i = 1, \ldots, 24)$, te $\widehat{\varepsilon}_1^2$ i $(\widehat{\varepsilon}_i - \widehat{\varepsilon}_{i-1})^2$.

Vrijednost $\widehat{d}$ test-statistike $\widehat{D}$ bit će

$$\widehat{d} = \frac{143{,}45}{102{,}4} \approx 1{,}4 \,,$$

pa se zaključuje, da uz razinu značajnosti $\alpha = 0{,}01$, hipotezu H_0 (greške mjerenja su nekorelirane slučajne varijable) treba prihvatiti. Iz tabl. XII. vidi se, naime, da su za $\alpha = 0{,}01$, $n = 24$ i $r = 1$ odgovarajući $c_1 = 1{,}04$, $d_1 = 1{,}20$, $d_2 = 2{,}80$ i $c_2 = 2{,}96$, pa $d = 1{,}4$ pada u područje prihvaćanja hipoteze H_0.

Uz $\alpha = 0{,}05$ dobili bismo $c_1 = 1{,}27$, $d_1 = 1{,}45$, $d_2 = 2{,}55$ i $c_2 = 2{,}73$, pa se tada ne bi mogla donijeti odluka o hipotezi H_0.



Slika 40. Skica odnosa veličina c_1, c_2, d_1, d_2 i $\widehat{d}$

Primjedba

Sve ono što je već rečeno o primjeni računala, posebno u vezi s regresijskom analizom, vrijedi i za analizu varijance. Na tržištu softvera postoje vrlo sofisticirani statistički programski paketi, koji omogućuju lako i brzo rješavanje i vrlo složenih zadataka.

Tablica 12.

i	y_i	$\widehat{\mu}_i$	$\widehat{\varepsilon}_i$	$\widehat{\varepsilon}_i^2$	$(\widehat{\varepsilon}_i - \widehat{\varepsilon}_{i-1})^2$
1	20	24,5	−4,5	20,25	
2	22		−2,5	6,25	4
3	24		−0,5	0,25	4
4	26		1,5	2,25	4
5	28	30,8	−2,8	7,84	18,49
6	30		−0,8	0,64	4
7	32		1,2	1,44	4
8	34		3,2	10,24	4
9	36	37,1	−1,1	1,21	18,49
10	38		0,9	0,81	4
11	40		2,9	8,41	4
12	40		2,9	8,41	0
13	43	43,4	−0,4	0,16	10,89
14	45		1,6	2,56	4
15	45		1,6	2,56	0
16	46		2,6	6,76	1
17	48	49,7	−1,7	2,89	18,49
18	50		0,3	0,09	4
19	50		0,3	0,09	0
20	52		2,3	5,29	4
21	53	56	−3	9	28,09
22	54		−2	4	1
23	55		−1	1	1
24	56		0	0	1
			$\sum$	102,4	143,45

Zadaci

1. Dokažite da za veličine δ_i $(i = 1, \ldots, m)$ definirane u (5) vrijedi $\sum_{i=1}^{m} n_i \delta_i = 0$.

2. Dokažite jednakost (13).

3. Dokažite da, bez obzira na hipotezu H_0 iz (7), za statistiku:

a) $\overline{Y}_i$ iz (8) vrijedi $\overline{Y}_i \sim N\left(\mu + \delta_i, \frac{1}{n_i}\sigma^2\right)$,

b) $\overline{Y}$ iz (9) vrijedi $\overline{Y} \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$.

4. Dokažite formule (22).
Uputa: Pozovite se na rezultate iz zad. 3. i činjenicu da se može pisati $Q_1 = \sum_{i=1}^{m} n_i \overline{Y}_i^2 - n\overline{Y}^2$.

5. Dokažite valjanost jednadžbi (26).

6. Dokažite da za statistiku:

a) $\overline{Y}$ iz (32) vrijedi $\overline{Y} \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$,

b) $\overline{Y}'_i$ iz (33) vrijedi $\overline{Y}'_i \sim N\left(\mu + \delta'_i, \frac{1}{m_2}\sigma^2\right)$,

c) $\overline{Y}''_j$ iz (34) vrijedi $\overline{Y}''_j \sim N\left(\mu + \delta''_j, \frac{1}{m_1}\sigma^2\right)$.

Uputa: Pozovite se na (31) i (26).

7. Dokažite valjanost jednakosti (39).
Uputa: Pođite od identiteta $(Y_{ij} - \overline{Y})^2 = [(Y_{ij} - \overline{Y}'_i - \overline{Y}''_j + \overline{Y}) + (\overline{Y}'_i - \overline{Y}) + (\overline{Y}''_j - \overline{Y})]^2$.

8. Dokažite jednakost (40).
Uputa: Iskoristite rezultate zad. 6. i činjenicu da se može pisati $Q_1 = m_2 \sum_{i=1}^{m_1} (\overline{Y}'_i)^2 - m_1 \overline{Y}^2$.

9. Dokažite jednakost (41).
Uputa: Vidite uputu za zad. 8.

10. Dokažite da za statistiku Q iz (35) vrijedi

$$E[Q] = (n-1)\sigma^2 + m_2 \sum_{i=1}^{m_1} (\delta'_i)^2 + m_1 \sum_{j=1}^{m_2} (\delta''_j)^2, \quad n = m_1 m_2 .$$

Uputa: Iskoristite rezultate zad. 6, činjenicu da se može pisati $Q = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} Y_{ij}^2 - n\overline{Y}^2$, te (26) i (31).

11. Dokažite jednakost (42).
Uputa: Iskoristite rezultat zad. 10, (39), (40) i (41).

12. Dokažite valjanost jednadžbi (57).

13. Dokažite da za statistiku:

a) $\overline{Y}$ iz (58) vrijedi $\overline{Y} \sim N\left(\mu, \frac{1}{n}\sigma^2\right)$, $n = m_1 m_2 l$,

b) $\overline{Y}'_i$ iz (59) vrijedi $\overline{Y}'_i \sim N\left(\mu + \delta'_i, \frac{1}{m_2 l}\sigma^2\right)$,

c) $\overline{Y}''_j$ iz (60) vrijedi $\overline{Y}''_j \sim N\left(\mu + \delta''_j, \frac{1}{m_1 l}\sigma^2\right)$,

d) $\overline{Y}_{ij}$ iz (61) vrijedi $\overline{Y}_{ij} \sim N\left(\mu + \delta'_i + \delta''_j + \delta_{ij}, \frac{1}{l}\sigma^2\right)$.

Uputa: Pozovite se na relacije (53)-(57).

14. Dokažite jednakost (67).
Uputa: Pođite od identiteta

$$(Y_{ijk} - \overline{Y})^2 = [(Y_{ijk} - \overline{Y}'_i - \overline{Y}''_j + \overline{Y}) + (\overline{Y}'_i - \overline{Y}) + (\overline{Y}''_j - \overline{Y})]^2 .$$

15. Dokažite da za statistiku Q iz (62) vrijedi

$$E[Q] = (n-1)\sigma^2 + m_2 l \sum_{i=1}^{m_1} (\delta'_i)^2 + m_1 l \sum_{j=1}^{m_2} (\delta''_j)^2 + 2l \left[\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \delta_{ij}(\delta'_i + \delta''_j + \delta_{ij})\right],$$

$n = m_1 m_2 l$.

16. Dokažite jednakost (68).
Uputa: Iskoristite rezultate zad. 13. i činjenicu da je

$$Q_1 = m_2 l \left[\sum_{i=1}^{m_1} (\overline{Y}'_i)^2 - m_1 \overline{Y}^2\right] .$$

17. Dokažite jednakost (69).
Uputa: Vidite uputu za zad. 16.

18. Dokažite jednakost (70)
Uputa: Iskoristite rezultat zad. 15, (67), (68), (69) i (71).

19. Dokažite jednakost (71).
Uputa: Vidite uputu za zad. 16.

20. Da bi se ispitao utjecaj vrste hrane na prirast težine tovljenika uzete su tri hranjive smjese A, B i C. Smjesom A hranjena su 4 tovljenika, smjesom B njih 3 i smjesom C 5 tovljenika. Mjerenjem težine tovljenika nakon mjesec dana ustanovljen je prirast (u postocima), što je prikazano ovom tablicom:

Vrsta hrane	Prirast težine				
A	13,7	14,2	12,8	13,7	
B	14,0	13,9	11,7		
C	13,7	14,2	13,3	14,0	14,1

a) Načinite odgovarajuću ANOVA-tablicu.
b) Može li se smatrati da vrsta hrane ne utječe na prirast težine?

21. Na svakoj od 12 jednakih parcela primijenjena je jedna od četiri vrste sjemena pšenice i jedna od tri vrste umjetnih gnojiva. Nakon žetve izmjereni su dobiveni prinosi (u tonama), što je prikazano ovom tablicom:

Vrsta umjetnog gnojiva \ Vrsta sjemena	1.	2.	3.	4.
1.	7,7	8,6	8,8	8,4
2.	8,1	9,2	9,1	9,3
3.	8,3	7,5	8,0	7,9

a) Načinite odgovarajuću ANOVA-tablicu.
b) Testirajte, uz razinu značajnosti $\alpha = 0{,}01$, hipotezu da vrsta sjemena ne utječe na prinos.
c) Testirajte hipotezu da vrsta umjetnog gnojiva ne utječe na prinos.

Uputa: Primijenite dvofaktorski aditivni model.

22. Na temelju rezultata iz tabl. 8. testirajte, uz razinu značajnosti $\alpha = 0{,}05$, hipotezu:

a) da marka automobila ne utječe na potrošnju goriva,
b) da vozačko iskustvo ne utječe na potrošnju goriva,
c) da nema interakcijskog djelovanja navedenih faktora na potrošnju goriva.

23. Svaka parcela iz zad. 21. sastoji se od dva jednaka dijela, pa je izmjeren prinos na svakom dijelu posebno, što je prikazano ovom tablicom:

Vrsta umjetnog gnojiva \ Vrsta sjemena	1.	2.	3.	4.
1.	3,5 4,2	4,2 4,4	4,0 4,8	3,9 4,5
2.	4,1 4,0	4,5 4,7	4,1 5,0	4,8 4,5
3.	4,0 4,3	3,5 4,0	4,2 3,8	3,6 4,3

a) Načinite odgovarajuću ANOVA-tablicu.
b) Uz razinu značajnosti $\alpha = 0{,}05$ testirajte hipotezu da vrsta sjemena ne utječe na prinos, da vrsta gnojiva ne utječe na prinos i da nema interakcijskog djelovanja na prinos.

XIV. Neparametarske metode

1. Uvod u problematiku

Za neparametarske metode teorije statističkog zaključivanja karakteristično je da se pri izgradnji odgovarajućeg matematičkog modela ne ističe pretpostavka o tipu razdiobe vjerojatnosti, tako da su te metode prikladne za one probleme u kojima se ne poznaje tip razdiobe vjerojatnosti iz koje potječu dani statistički podaci. Vidjeli smo, naime, da je većina razmotrenih modela sadržavala pretpostavku da je riječ o klasi normalnih razdioba $N(\mu, \sigma^2)$, pri čemu je nepoznat jedan parametar, ili oba parametra μ i σ^2. U nekim modelima pretpostavljali su se i drugi jednoparametarski i višeparametarski tipovi vjerojatnosnih razdioba (Poissonova, eksponencijalna, binomna, uniformna, dvodimenzionalna normalna i sl.).

U neparametarskim modelima obično se pretpostavlja da se klasa dopuštenih razdioba vjerojatnosti sastoji od svih kontinuiranih vjerojatnosnih razdioba. Ona je, dakako, mnogo opsežnija od ranije promatranih klasa dopuštenih razdioba u parametarskim modelima, što upućuje na zaključak da će se neparametarskim metodama dobivati slabiji zaključci (manje efikasni procjenitelji, rizičnije odluke pri testiranju hipoteza i sl.) nego parametarskim metodama, posebno kada se raspolaže s malo podataka. Međutim, kada je broj podataka velik, taj nedostatak neparametarskih metoda iščezava, pa u tome i jest njihova praktična vrijednost.

Tipičan primjer primjene neparametarskih metoda pri procjeni parametara opisan je već u VI.2, gdje su se promatrali procjenitelji za očekivanje i varijancu, pri čemu se klasa dopuštenih razdioba vjerojatnosti sastojala od svih razdioba konačne varijance, odnosno konačnoga četvrtog centralnog momenta. Dobiveni procjenitelji, posebno na velikim uzorcima ($n \to \infty$), imaju neka vrlo dobra svojstva (konzistentnost, asimptotska normalnost i dr.), ali je nezgoda što se malo može reći o njihovoj efikasnosti.

U neparametarskim modelima za procjenu parametara neprimjenjiva je metoda najmanje vjerojatnosti za dobivanje procjenitelja, čime smo lišeni svih onih pogodnosti koje imaju ML-procjenitelji (v. VI.6). Općenito se može reći da neparametarski modeli ne omogućuju primjenu općih načela (metoda najveće vjerojatnosti, metoda momenata) za dobivanje procjenitelja promatranih parametara, nego se obično procjenitelj definira na temelju intuitivnog uvida u odnose promatranog parametra i određenih statistika.

Budući da u neparametarskim modelima testiranja hipoteza nije moguće klasu dopuštenih razdioba vjerojatnosti karakterizirati pomoću konačnog broja parametara, pojavljuju se teškoće u preciznom definiranju nul-hipoteze i alternativne hipoteze. Pri testiranju tzv. *neparametarskih hipoteza*, odnosno pri konstrukciji *neparametarskih testova*, komplicira se pojam funkcije snage, odnosno operativne karakteristike testa. Test-statistika kod neparametarskih testova redovito se definira na temelju određenoga intuitivnog uvida u pokazatelje odstupanja od nul-

-hipoteze, a ne na temelju nekoga općeg načela, kao kod parametarskih testova. Općenito se smatra da je odbacivanje nul-hipoteze vrlo dobro teorijski utemeljeno, dok je prihvaćanje nul-hipoteze u neparametarskim testovima prilično slabo argumentirano.

Klasičan primjer neparametarskog testa jest test opisan u IX.5, gdje se kao nul-hipoteza uzima da niz sparenih mjerenja (x_i, y_i) $(i = 1, \ldots, n)$ potječe od nezavisnih slučajnih varijabli X i Y. Vrijednost pripadne test-statistike (formula (34) iz IX.5) indicira odstupanje od nezavisnosti i prevelika vrijednost upućuje na odbacivanje hipoteze o nezavisnosti. Dobije li se, pak, mala pozitivna vrijednost spomenute test-statistike, koja uvjetuje prihvaćanje hipoteze o nezavisnosti, teško se može naći teorijsko objašnjenje da su promatrane slučajne varijable X i Y zaista nezavisne.

Zbog već istaknute činjenice da se neparametarski testovi obično konstruiraju tako da se intuitivno nasluti prikladna test-statistika za određeni problem, a zatim se istraži pripadna joj razdioba vjerojatnosti, u literaturi su opisani brojni neparametarski testovi za različite probleme statističkog zaključivanja. U nastavku će se prikazati neki najpoznatiji i u praksi najčešće primjenjivani neparametarski testovi.

2. Procjena medijana i kvantila

Pojam medijana i kvantila p-tog $(0 < p < 1)$ reda, kao određenih terijskih pojmova (parametara) u vezi s kontinuiranim razdiobama vjerojatnosti, uvedeni su u IV.4, gdje su opisana i njihova osnovna svojstva.

U mnogim praktičnim problemima potrebno je procijeniti medijan M, a često je potrebno procijeniti i kvantil x_p, p-tog reda, pa se prirodno nameće zahtjev za definiranjem odgovarajućih procjenitelja, na temelju danoga slučajnog uzorka $(X_1, \ldots, X_n)$.

Ako je F f.r.v. određene kontinuirane razdiobe vjerojatnosti, onda je kvantil x_p definiran formulom

(1) $$F(x_p) = p\,, \qquad 0 < p < 1\,.$$

Za $p = 0{,}5$ imamo medijan $M = x_{0,5}$, tako da vrijedi

(2) $$F(M) = 0{,}5\,.$$

Problemu definiranja procjenitelja za kvantil x_p može se, dakako, pristupiti i pomoću određenoga parametarskog modela, tako da se specificira klasa dopuštenih razdioba vjerojatnosti $\mathcal{P} = \{P_t : t \in \Theta\}$, gdje je t parametar (može biti i vektorski), a Θ dani skup dopuštenih vjerojatnosti parametra t. Neka je $\hat{t}$ vrijednost procjenitelja $\hat{T}$ za nepoznati parametar t, pa se tada $F_{\hat{t}}(x)$ može uzeti kao procjena za nepoznatu vrijednost $F_t(x)$ $(x \in \mathbf{R})$ funkcije razdiobe vjerojatnosti promatrane slučajne varijable X. Rješavanjem jednadžbe $F_{\hat{t}}(x_p) = p$, po x_p, dobiva se

$$x_p = \hat{x}_p = F_{\hat{t}}^{-1}(p)\,,$$

pa možemo smatrati da je statistika

(3) $$\hat{X}_p = F_{\hat{T}}^{-1}(p)\,, \qquad 0 < p < 1\,,$$

određeni procjenitelj za nepoznati kvantil x_p promatrane kontinuirane razdiobe vjerojatnosti u danom parametarskom modelu.

Pretpostavi li se još da je $\hat{T}$ ML-procjenitelj za t, na temelju svojstva invarijantnosti (v. VI.6) proizlazi da je i $\hat{X}_p$ iz (3) ML-procjenitelj za x_p.

Uvid u funkciju rizika (srednju kvadratnu grešku) pri procjeni nepoznatog kvantila x_p, vrijednošću $\hat{x}_p$ ML-procjenitelja iz (3), može se približno dobiti primjenom formule (68) iz VI.6.

1. primjer

Pretpostavimo da podaci $x_1, \ldots, x_n$ potječu od normalne razdiobe $N(\mu, \sigma^2)$, gdje su $\mu \in \mathbf{R}$ i $\sigma > 0$ nepoznati parametri $(\boldsymbol{t} = (\mu, \sigma))$. Vrijednost $F_{\boldsymbol{t}}(x)$ funkcije razdiobe vjerojatnosti za normalnu razdiobu $N(\mu, \sigma^2)$ može se izraziti pomoću vrijednosti funkcije razdiobe vjerojatnosti standardne normalne razdiobe (v. IV.5), tako da se može pisati

$$F_{\boldsymbol{t}}(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)\,, \qquad x \in \mathbf{R}\,.$$

Rješavanjem jednadžbe $\Phi\left(\frac{x_p - \mu}{\sigma}\right) = p$, po x_p, dobiva se

(4) $$x_p = \sigma\,\Phi^{-1}(p) + \mu = \sigma z_p + \mu\,,$$

gdje je $z_p = \Phi^{-1}(p)$.

Uzme li se $\hat{\boldsymbol{T}} = (\overline{X}, \hat{\Sigma})$ ($\overline{X}$ je uzoračka aritmetička sredina, a $\hat{\Sigma}$ uzoračka standardna devijacija) kao procjenitelj za nepoznati vektorski parametar $\boldsymbol{t} = (\mu, \sigma)$, bit će

(5) $$\hat{X}_p = z_p\hat{\Sigma} + \overline{X}$$

ML-procjenitelj za kvantil p-tog reda u parametarskom modelu s klasom dopuštenih razdioba $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbf{R},\ \sigma > 0\}$. Pozivajući se na rezultate iz VI.6. lako se nalazi (v. zad. 1) da je

(6) $$\mathrm{E}[\hat{X}_p] \approx x_p - \frac{1}{n}\left(1 - \frac{1}{4n}\right) z_p\,\sigma\,,$$

iz čega se razabire da $\hat{X}_p$ nije nepristrani procjenitelj za x_p, već da približno ima pristranost $-\frac{1}{n}\left(1 - \frac{1}{4n}\right) z_p\sigma$.

Slično se izvodi (v. zad. 1) formula

(7) $$\mathrm{E}[(\hat{X}_p - x_p)^2] \approx \frac{\sigma^2}{n}\left(1 + \frac{1}{2}\,z_p\sigma\right)\,,$$

iz čega se razabire da je $\widehat{X}_p$, definirano formulom (5), kozistentan procjenitelj za x_p.

Zanimljivo je primijetiti, što se razabire iz (7), da se očekivana kvadratna greška povećava kada se p približava jedinici ($p \to 1 \Rightarrow z_p \to \infty$). To pokazuje da procjenitelj $\widehat{X}_p$, iz (5), postaje sve lošiji za tzv. *repne kvantile*.

Problem nalaženja razdiobe vjerojatnosti procjenitelja $\widehat{X}_p$ iz (3) redovito je vrlo kompliciran zadatak, što praktički otežava utvrđivanje efikasnosti i određivanje pripadne funkcije rizika, tako da su parametarski modeli neprikladni pri definiranju procjenitelja za nepoznate kvantile.

U neparametarskom modelu uvodi se pojam uređajne statistike. Ako se mjerenja $x_1, \dots, x_n$, koja potječu od kotinuirane slučajne varijable X, poredaju po veličini tako da vrijedi

(8) $$x'_1 \le x'_2 \le \dots \le x'_n ,$$

onda se $x'_i = y_i$ ($i = 1, \dots, n$) zove vrijednost i-te uređajne statistike Y_i. Prema tome, i-ta ***uređajna statistika*** Y_i ($i = 1, \dots, n$) slučajna je varijabla koja pokazuje i–tu po veličini vrijednost, kada se izvede n nezavisnih mjerenja promatrane slučajne varijable X. Očigledno je

$$Y_1 = \min(X_1, \dots, X_n) , \quad Y_n = \max(X_1, \dots, X_n) .$$

Statistika

(9) $$\widehat{X}_p = Y_{[np]+1} , \qquad 0 < p < 1 ,$$

gdje $[np]$ označuje najveći cijeli broj koji ne premašuje np ("*najveće cijelo*" od np), zove se ***uzorački kvantil*** p-tog ***reda***. Vrijednost

(10) $$\widehat{x}_p = y_{[np]+1} = x'_{[np]+1}$$

uzoračkog kvantila p-tog reda izračunava se, prema tome, tako da se u neopadajućem nizu danih podataka (8) uzme član s indeksom (rednim brojem) $[np] + 1$. Specijalno za $p = 0{,}5$ dobiva se ***uzorački medijan*** $\widehat{M}$ i njegova je vrijednost na danom n-članom nizu podataka

(11) $$\widehat{m} = \widehat{x}_{0,5} = x'_{[\frac{n}{2}]+1} = \begin{cases} x'_{\frac{n+1}{2}} , & \text{za neparno n} \\ x'_{\frac{n}{2}+1} , & \text{za parno n} . \end{cases}$$

Sada se nameće pitanje o svojstvima statistike $\widehat{X}_p$ iz (9), kao procjenitelja za nepoznati kvantil x_p. Najzanimljiviji rezultat (v. [4]), koji se izvodi uz nešto složenije postupke, sastoji se u tome da za velike n približno vrijedi

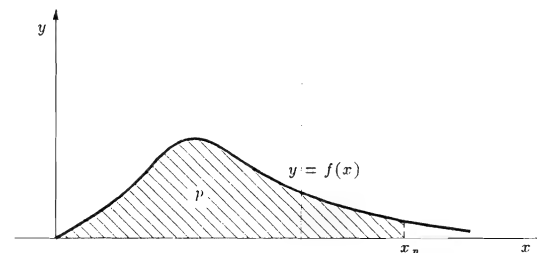
(12) $$\mathrm{E}[(\widehat{X}_p - x_p)^2] \approx \frac{1}{n} \frac{p(1-p)}{[f(x_p)]^2} , \qquad x_p \in \Theta ,$$

gdje je $\Theta = \{x \in \mathbf{R} : f(x) > 0\}$, a f je pripadna f.g.v.

Važna je činjenica, koja proizlazi neposredno iz (12), da je uzorački kvantil $\widehat{X}_p$ konzistentan procjenitelj za teorijski kvantil p-tog reda pripadne kontinuirane razdiobe vjerojatnosti. To opravdava postupak da se za velike n nepoznati kvantil x_p aproksimira vrijednošću $\widehat{x}_p$ procjenitelja $\widehat{X}_p$ iz (9).

Iz (12) se također razabire da će $\widehat{X}_p$ iz (9) biti slab procjenitelj za repne kvantile određenih vjerojatnosnih razdioba. Ako je, naime, p blizu jedinice, onda je u mnogim važnim vjerojatnosnim razdiobama (normalna razdioba, γ-razdioba, F-razdioba i sl.) x_p vrlo veliko, a $f(x_p)$ vrlo maleno ($p \to 1 \Rightarrow x_p \to \infty \Rightarrow f(x_p) \to \to 0$), pa (12) pokazuje da će očekivana kvadratna greška, u tom slučaju biti vrlo velika.

Prema tome, procjena repnih kvantila, u takvim situacijama, pomoću procjenitelja definiranih formulama (3) i (9), ima smisla samo onda ako je broj podataka (n) vrlo velik (v. zad. 5).



Slika 41. Grafički prikaz repnog kvantila

3. Intervali povjerenja za kvantile

Da bi se egzaktnije uočila veličina greške pri procjeni nepoznatog kvantila x_p, razmotrit će se problem određivanja intervala povjerenja zadane pouzdanosti γ u neparametarskom modelu. Ako su vrijednosti u nizu izmjerenih podataka poredane po veličini, kao u (8), i ako je vrijednost procjene $\widehat{x}_p$ definirana kao u (10), onda se prirodno nameće ideja da se kao vrijednost rubova intervala povjerenja za nepoznati parametar x_p uzmu

(13) $$g_1 = x'_{r-k} , \qquad g_2 = x'_{r+k} ,$$

gdje je $r = [np] + 1$, a $k \in \mathbf{N}$ odrediti će se tako da interval povjerenja $[G_1, G_2]$ ima zadanu pouzdanost $\gamma (0 < \gamma < 1)$. Prema tome, lijevi je rub g_1 intervala povjerenja za x_p vrijednost $(r-k)$-te, a desni vrijednost $(r+k)$-te uređajne statistike. Stavi li se $r-k = i$, a $r+k = j$, zadatak se može formulirati tako da se zahtijeva određivanje prirodnih brojeva i i $j (i < j)$ takvih da slučajni interval $[Y_i, Y_j]$ pokrije nepoznati kvantil x_p bar s vjerojatnošću γ.

Budući de je x'_i vrijednost slučajne varijable Y_i, koja označuje i-tu po veličini izmjerenu vrijednost u danom n-članom slučajnom uzorku, a x'_j vrijednost slučajne varijable Y_j, koja označuje j-tu po veličini izmjerenu vrijednost, riječ je o tome da se za fiksirane n (veličina uzorka), p (red kvantila) i γ (pouzdanost) odrede prirodni brojevi i i j ($i < j$), tako da vrijedi

(14) $$\mathrm{P}(Y_i \le x_p \le Y_j) \ge \gamma ,$$

pri čemu će se, naravno, težiti da i i j budu međusobno što bliži.

Temeljna činjenica koja omogućuje rješenje postavljenog zadatka sadržana je u formuli

(15)	$P(Y_i \le x_p \le Y_j) = P(i \le Z_n < j)\,, \qquad Z_n \sim B(n,p)\,.$

Formula (15) pokazuje da je vjerojatnost da interval $[Y_i, Y_j]$, sa slučajnim rubovima $G_1 = Y_i$ i $G_2 = Y_j$, pokrije kvantil x_p, nepoznate kontinuirane razdiobe vjerojatnosti, jednaka vjerojatnosti da binomna slučajna varijabla Z_n poprimi vrijednost koja nije manja od i, a manja je od j.

Da bi se dokazala ispravnost formule (15) rezonira se ovako: Nejednakošću $Y_i \le x_p \le Y_j$ opisan je slučajni događaj koji je ekvivalentan simultanom nastupanju slučajnih događaja

(16) $$\{Y_i \le x_p\} \quad \text{i} \quad \{Y_j \ge x_p\}\,.$$

Neka $A = \{X \le x_p\}$ označuje događaj da je pri mjerenju slučajne varijable X dobivena vrijednost koja nije veća od x_p, a Z_n neka označuje broj nastupa događaja A prilikom n nezavisnih ponavljanja toga mjerenja. Može se reći da Z_n označuje broj onih vrijednosti (rezultata mjerenja) u n-članome slučajnom uzorku koje nisu veće od x_p. Budući da događaj $\{Y_i \le x_p\}$ iskazuje da i-ta uređajna statistika nije veća od x_p, taj je događaj ekvivalentan događaju $\{Z_n \ge i\}$. Događaj $\{Y_j \ge x_p\}$ iskazuje da j-ta uređajna statistika nije manja od x_p, pa je tom događaju ekvivalentan događaj $\{Z_n < j\}$. Očigledno je da se nejednakosti $Z_n \ge i$ i $Z_n < j$ mogu zapisati i kao $i \le Z_n < j$, čime je dokazano da su $\{Y_i \le x_p \le Y_j\}$ i $\{i \le Z_n < j\}$ ekvivalentni slučajni događaji.

Vjerojatnost događaja A je $P(A) = P(X \le x_p) = F(x_p) = p$. Budući da slučajna varijabla Z_n označuje broj pojavljivanja događaja A, vjerojatnosti p, prilikom n nezavisnih ponavljanja slučajnoga eksperimenta (mjerenja slučajne varijable X), prema onome što je rečeno u IV.3. slučajnoj varijabli Z_n pripada binomna razdioba s parametrima n i p. Time je dokazana tvrdnja formule (15), pa to omogućuje da se napiše

(17) $$P(G_1 \le x_p \le G_2) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{k-1} = \gamma\,,$$

što pokazuje da intervalu, sa slučajnim rubovima $G_1 = Y_i$ i $G_2 = Y_j$, pripada pouzdanost γ, kao intervalu povjerenja za nepoznati kvantil p-tog reda.

Prema tome, da bi se odredio interval povjerenja unaprijed zadane pouzdanosti γ za nepoznati kvantil p-tog reda neke kontinuirane razdiobe vjerojatnosti, na temelju n-članog slučajnog uzorka, treba najprije odrediti cijele brojeve i i j tako da bude zadovoljena nejednakost

(18) $$P(i \le Z_n < j) \ge \gamma\,, \qquad Z_n \sim B(n,p)\,,$$

uz najmanju moguću razliku $j - i$. Zatim se i-ta po veličini vrijednost $x'_i = g_1$ uzima kao donji, a j-ta vrijednost $x'_j = g_2$ iz uređenoga slučajnog uzorka kao gornji rub intervala povjerenja $[g_1, g_2]$ pouzdanosti γ.

Za određivanje vrijednosti i i j u konkretnim primjerima mogu se primijeniti tablice binomne razdiobe (v. tabl. I. u Dodatku).

U tabl. 1. navedene su vrijednosti za i i j pri procjeni nepoznatog medijana $M = x_{0,5}$ uz pouzdanost $\gamma = 0{,}95$, u ovisnosti o veličini uzorka n.

Tablica 1.

n	10	20	30	40	50
i	2	6	10	14	18
j	9	15	21	27	33

Tako, na primjer, ako je $n = 30$, onda za donji rub intervala povjerenja pouzdanosti 95 % treba uzeti desetu, a za gornji rub dvadeset prvu po veličini vrijednost iz uređenog niza od 30 mjerenja promatrane slučajne varijable.

Zanimljivo je primijetiti da je, prema (11), u ovom slučaju $\hat{m} = x'_{16}$, tj. vrijednost neparametarskog procjenitelja za nepoznati medijan M ne nalazi se nužno u sredini pripadnog intervala povjerenja.

Za velike n postupak određivanja rubova intervala povjerenja može se pojednostavniti, uz primjenu činjenice da se tada binomna razdioba $B(n,p)$ može aproksimirati normalnom razdiobom $N(np, np(1-p))$ (v. zad. 27. u VI. poglavlju). Stavi li se, naime, u (18) $i = np - k$ i $j = np + k$, može se pisati

$$P(np - k \le Z_n < np + k) \ge \gamma\,,$$

pri čemu se uzima da približno vrijedi $Z_n \sim N(np, np(1-p))$, iz čega proizlazi

$$2\Phi\left(\frac{k}{\sqrt{np(1-p)}}\right) - 1 \ge \gamma\,,$$

odnosno

(19) $$k \ge k_0 = \sqrt{np(1-p)}\ \Phi^{-1}\left(\frac{\gamma+1}{2}\right) = z_\gamma \sqrt{np(1-p)}\,,$$

gdje je $z_\gamma = \Phi^{-1}\left(\frac{\gamma+1}{2}\right)$. Za uobičajene vrijednosti γ veličina z_γ prikazana je u tabl. 1. u VII.1.

Prema tome, interval povjerenja pouzdanosti γ za nepoznati kvantil p-tog reda približno ima rubove

(20) $$g_1 = x'_{i_0}\,, \qquad g_2 = x'_{j_0}\,,$$

gdje je i_0 cijeli broj najbliži broju $np - k_0$, a j_0 cijeli broj najbliži broju $np + k_0$.

Ako se, na primjer, želi odrediti interval povjerenja pouzdanosti $\gamma = 0{,}95$ za medijan M, na temelju $n = 100$ mjerenja promatrane kontinuirane slučajne varijable, onda je $p = 0{,}5$ i $z_\gamma = z_{0,95} = 1{,}96$, dok je $\sqrt{np(1-p)} = \sqrt{25} = 5$, tako da je $k_0 = 9{,}8$, iz čega slijedi da je $g_1 = x'_{40}$ i $g_2 = x'_{60}$, tj. lijevi je rub četrdeseta, a desni rub intervala povjerenja pouzdanosti 95 % za medijan šezdeseta po veličini vrijednost u danom nizu od 100 mjerenja promatrane kontinuirane slučajne varijable.

Žele li se odrediti intervali povjerenja pouzdanosti 95 % za kvartile $x_{0,25}$ i $x_{0,75}$ promatrane kontinuirane razdiobe vjerojatnosti (v. IV.4), uzet će se najprije $p = 0{,}25$ i prema (19) izračunati pripadni $k_0 = z_{0,95}\sqrt{100 \cdot 0{,}25 \cdot 0{,}75} \approx \approx 1{,}96 \cdot 4{,}33 \approx 8{,}49$, pa iz (20) dobivamo

$$g_1 = x'_{17}\,, \qquad g_2 = x'_{33}\,,$$

što znači da se sedamnaesta vrijednost u uređenom nizu od 100 podataka uzima kao donji rub, a trideset treća vrijednost kao gornji rub intervala povjerenja pouzdanosti 95 % za nepoznati lijevi kvartil $x_{0,25}$.

Na sličan bi način našli i granice intervala povjerenja za desni kvartil $x_{0,75}$ (v. zad. 7).

Primjenom formule (10) dobivaju se, inače, točkaste procjene za kvartile, što bi u ovom slučaju dalo

$$\widehat{x}_{0,25} = y_{26} = x'_{26}\,, \qquad \widehat{x}_{0,75} = y_{76} = x'_{76}\,,$$

što pokazuje da je dvadeset šesta vrijednost u uređenom nizu od 100 podataka procjena za lijevi kvartil, a sedamdeset šesta za desni kvartil.

4. Test predznaka

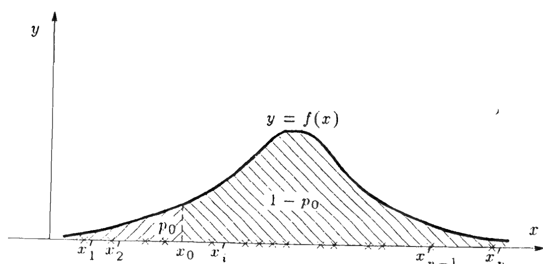
U vezi s kvantilima kontinuirane razdiobe vjerojatnosti logično se odmah postavlja i zadatak testiranja hipoteza o kvantilima. Neka F označuje f.r.v. kontinuirane razdiobe vjerojatnosti iz koje potječu dani podaci $x_1, \ldots, x_n$ i neka su $x_0 (x_0 \in \mathbf{R})$ i p_0 $(0 < p_0 < 1)$ zadani brojevi. Tada se može postaviti zadatak konstrukcije testa za testiranje nul-hipoteze $\mathrm{H}_0 : x_{p_0} = x_0$, prema nekoj od alternativnih hipoteza ($\mathrm{H}_1 : x_{p_0} \neq x_0$, $\mathrm{H}_1 : x_{p_0} < x_0$, $\mathrm{H}_1 : x_{p_0} > x_0$). Nul-hipotezom ističe se slutnja da kvantil p_0-tog reda ima vrijednost x_0. Ako je to istina, onda vrijedi jednakost

$$F(x_0) = p_0\,, \tag{21}$$

i obratno. Možemo, dakle, nul-hipotezu zapisati i u obliku

$$\mathrm{H}_0 : F(x_0) = p_0\,, \tag{22}$$

pa se problem testiranja hipoteze H_0 može shvatiti i kao provjera, pomoću danih podataka, slutnje da intervalu $\langle -\infty, x_0]$ pripada vjerojatnost p_0, a intervalu $[x_0, \infty\rangle$ vjerojatnosti $1 - p_0$.



Slika 42. Skica problema testiranja hipoteze o kvantilu

Sjetimo li se Pearsonova teorema (v. IX.1), odmah nam se nameće ideja da se zadatak formulira kao primjena hikvadrat-testa ($r = 2$), tj. kao testiranje nul-hipoteze

$$\mathrm{H}_0 : (p_1 = p_0\,,\ p_2 = 1 - p_0)\,, \tag{23}$$

pomoću test-statistike (v. (7) u IX.1)

$$D = \frac{(\widehat{F}_1 - f_1^{(0)})^2}{f_1^{(0)}} + \frac{(\widehat{F}_2 - f_2^{(0)})^2}{f_2^{(0)}} \sim \chi^2(1)\,, \tag{24}$$

gdje su $f_1^{(0)} = np_0$ i $f_2^{(0)} = n(1 - p_0)$ teorijske frekvencije, a $\widehat{F}_1$ i $\widehat{F}_2 = n - \widehat{F}_1$ empirijske frekvencije, tj. broj podataka n-članog niza statističkih podataka u intervalu $\langle -\infty, x_0]$, odnosno u intervalu $\langle x_0, \infty\rangle$. Dobije li se prevelika vrijednost d, test-statistike D iz (24), hipoteza H_0 će se odbaciti. Uzme li se razina značajnosti α, hipoteza H_0 se odbacuje pri

$$d \geq \mathrm{H}_1^{-1}(1 - \alpha)\,, \tag{25}$$

gdje je H_1 f.r.v. za hikvadrat-razdiobu s jednim stupnjem slobode (v. tabl. VI. u Dodatku).

Poznato je da Pearsonov teorem zahtijeva veliki broj podataka ($n \to \infty$), pa se prirodno postavlja zadatak nalaženja testa koji neće imati navedeno ograničenje. To se zaista lako postiže, ako se kao test-statistika uzme slučajna varijabla Z_n, koja označuje broj članova u n-članom nizu podataka koji nisu veći od x_0, odnosno broj nenegativnih članova u nizu $x_0 - x_1, x_0 - x_2, \ldots, x_0 - x_n$. Očigledno je da u uvjetima istinitosti hipoteze H_0 vrijedi

$$Z_n \sim \mathrm{B}(n, p_0)\,. \tag{26}$$

Međutim, bez obzira na istinitost hipoteze H_0, slučajna varijabla $Z_n \sim \mathrm{B}(n,p)$ $(0 < p < 1)$, pa se za Z_n očekuje vrijednost $\mathrm{E}[Z_n] = np$. Ako hipoteza H_0 stvarno nije istinita i ako je, recimo, x_{p_0} mnogo veće od x_0, tj. ako je $p = F(x_0)$ mnogo manje od p_0, onda se može očekivati mnogo manja vrijednost test-statistike Z_n od np_0.

Postavi li se zadatak da se testira $\mathrm{H}_0 : x_{p_0} = x_0$, prema alternativnoj hipotezi $\mathrm{H}_1 : x_{p_0} > x_0$, uz razinu značajnosti α, kritično će područje biti oblika

$$C_1 = \langle -\infty, c_1]\,, \tag{27}$$

gdje je c_1 određeno tako da vrijedi

$$\sum_{k=0}^{c_1} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = \alpha\,. \tag{28}$$

Dobije li se, na danim podacima, vrijednost z_n test-statistike Z_n iz skupa C_1, to pokazuje da treba posumnjati u nul-hipotezu. U uvjetima istinitosti hipoteze H_0 očekuje se, naime, da broj podataka koji nisu veći od x_{p_0}, tj. broj "uspjeha" u Bernoullijevu nizu pokusa uz vjerojatnost p_0, neće biti premalen (manji od kritične vrijednosti c_1 određene zahtjevom (28)).

Uzme li se hipoteza $H_1 : x_{p_0} < x_0$, kao alternativna hipoteza, ona se može zapisati i kao $H_1 : F(x_0) > p_0$, pa će kritično područje, očigledno, imati oblik

$$C_2 = [c_2, \infty\rangle \,, \tag{29}$$

gdje je c_2 određeno tako da vrijedi

$$\sum_{k=c_2}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha \,. \tag{30}$$

Prevelika vrijednost z_n test-statistike Z_n iz (26) upućuje na odbacivanje nul-hipoteze da je kvantil p_0-tog reda jednak broju x_0 i sugerira prihvaćanje alternativne hipoteze da je kvantil p_0-tog reda manji od broja x_0.

Uzimajući $H_1 : x_{p_0} \neq x_0$ kao alternativnu hipotezu, kritično će područje biti oblika

$$C_0 = \langle -\infty, c_1] \cup [c_2, \infty\rangle \,, \tag{31}$$

gdje su c_1 i c_2 ($c_1 < c_2$) određeni tako da vrijedi

$$\sum_{k=0}^{c_1} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \sum_{k=c_2}^{n} \binom{n}{k} p_0^k (1-p_0)^k = \frac{\alpha}{2} \,. \tag{32}$$

Primjedba

Očigledno je da, zbog diskretnosti binomne razdiobe, neće uvijek biti moguće naći točno rješenje jednadžbi (28), (30) i (32), pa će se zato uzeti one cjelobrojne vrijednosti za c_1 i c_2 koje daju najbliže vrijednosti odgovarajućih zbrojeva zadanoj razini značajnosti α.

2. primjer

Neka je $n = 10$ i $p_0 = 0{,}5$, pa treba odrediti kritično područje razine značajnosti $\alpha = 0{,}10$ za testiranje hipoteze $H_0 : M = 0$ ($M = x_{0,5}$), prema alternativnoj hipotezi $H_1 : M \neq 0$. Da bismo odredili c_1 i c_2 iz (32), primijetimo da je za $p_0 = 0{,}5$ binomna razdioba $B(10\,;0{,}5)$ simetrična razdioba, tako da je dovoljno naći c_1, jer je tada $c_2 = n - c_1$.

Pogleda li se tabl. I. u Dodatku, vidi se da je

$$\sum_{k=0}^{2} \binom{10}{k} 0{,}5^k \cdot 0{,}5^{10-k} = \sum_{k=8}^{10} \binom{10}{k} 0{,}5^{10} = 0{,}0547 \,,$$

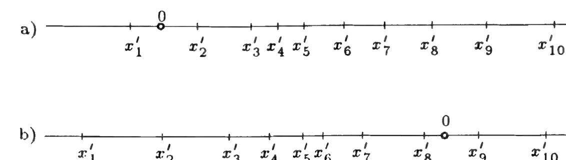
tako da kritičnom području oblika (31), uz $c_1 = 2$ i $c_2 = 10 - 2 = 8$, pripada razina značajnosti $2 \cdot 0{,}0547 = 0{,}1094$ i to je najbliže traženoj razini značajnosti $\alpha = 0{,}10$.

Da smo uzeli $c_1 = 1$ i $c_2 = 9$, dobili bismo

$$\sum_{k=0}^{1} \binom{10}{k} 0{,}5^{10} = 0{,}0107$$

i tada bi kritičnom području $\langle -\infty, 1] \cup [9, \infty\rangle$ pripadala razina značajnosti $2 \cdot 0{,}0107 = 0{,}0214$, što se mnogo više razlikuje od 0,10.

Prema tome, ako se na 10-članome slučajnom uzorku dobije dva ili manje (sl. 43a), odnosno osam ili više (sl. 43b) nepozitivnih podataka (mjerenja promatrane slučajne varijable), hipoteza da je medijan te slučajne varijable nula će se odbaciti, uz rizik od 10 % da se može odbaciti i stvarno istinita hipoteza.



Slika 43. Skica situacija u kojima se odbacuje nul-hipoteza

Najčešća je primjena testa predznaka baš za $p = 0{,}5$ i $x_0 = 0$, tj. testiranje hipoteze o tome da je medijan nula, prema nekoj od spomenutih alternativnih hipoteza. Osim toga, niz statističkih podataka $x_1, \ldots, x_n$ obično je dobiven kao $x_i = u_i - v_i$, pri čemu se pretpostavlja da uređeni parovi (u_i, v_i) $(i = 1, \ldots, n)$ potječu od nezavisnih kontinuiranih slučajnih varijabli U_i i V_i sa zajedničkom razdiobom vjerojatnosti. U tom je, naime, slučaju $X_i = U_i - V_i$ kontinuirana slučajna varijabla kojoj pripada simtrična oko nule razdioba vjerojatnosti, čiji je medijan, naravno, nula.

Stoga se zadatak o provjeri slutnje da spareni podaci (u_i, v_i) $(i = 1, \ldots, n)$ potječu od nezavisnih slučajnih varijabli U_i i V_i, kojima pripada ista kontinuirana razdioba vjerojatnosti, formulira kao testiranje hipoteze da niz $x_i = u_i - v_i$ $(i = 1, \ldots, n)$ potječe od vjerojatnosne razdiobe s medijanom nula.

Pripomenimo da se ne pretpostavlja ista razdioba vjerojatnosti za slučajne varijable $U_1, \ldots, U_n$, odnosno $V_1, \ldots, V_n$.

3. primjer

Da bi se provjerila hipoteza o jednakoj kvaliteti auto-guma marke A i marke B, mjeri se broj prijeđenih kilometara do istrošenosti auto-gume. Izabrano je 10 različitih vozila na koja su najprije stavljene gume marke A, a nakon istrošenosti gume marke B. Zabilježeni su ovi rezultati (u stotinama kilometara):

Tablica 2.

u_i (marka A)	301	197	97	188	252	341	376	315	293	168
v_i (marka B)	282	184	89	203	289	312	358	333	279	142
$\operatorname{sign}(x_i)$	+	+	+	−	−	+	+	−	+	+

Apstraktno-matematički gledano, zadatak je identičan onome u 2. primjeru, jer je riječ o tome da se testira hipoteza $H_0 : M = 0$, prema alternativnoj hipotezi $H_1 : M \neq 0$, na temelju danog niza od 10 podataka iz trećeg retka tabl. 2. Pitamo

se, zapravo, smije li se zaključiti da se gume marke A i marke B značajno razlikuju po kvaliteti ako se od deset izvršenih mjerenja u tri slučaja pokazalo da je auto-guma marke A lošija od auto-gume marke B.

U 2. primjeru izveli smo da kritičnom području $\langle -\infty, 1] \cup [9, \infty\rangle$ pripada razina značajnosti 2,14 %, a kritičnom području $\langle -\infty, 2] \cup [8, \infty\rangle$ pripada razina značajnosti 10,94 %. Za kritično područje $\langle -\infty, 3] \cup [7, \infty\rangle$ nalazi se da je pripadna razina značajnosti 34,38 %.

Budući da je vrijednost test-statistike $z_{10} = 3$, vidi se da, uz uobičajene razine značajnosti (5 % ili 10 %), nul-hipotezu (ne postoji značajna razlika u kvaliteti auto-guma marke A i B) treba prihvatiti.

Dobra su svojstva razmotrenog testa, tzv. ***testa predznaka***, da je vrlo jednostavan za primjenu, jer praktički ne zahtijeva nikakva računanja, a valjan je uz minimalne teorijske pretpostavke. Očigledno je, međutim, da se u testu predznaka ne upotrebljava velik dio informacije sadržane u danim statističkim podacima, jer se iskorištava samo predznak razlike $u_i - v_i$, a ne i apsolutna vrijednost. U podacima mora postojati jako izražena tendencija "pomicanja udesno, ili ulijevo" da bi se donijela odluka o odbacivanju nul-hipoteze, uz uobičajene razine značajnosti. Moramo konstatirati da se nul-hipoteza vrlo olako prihvaća.

Izračunavanje veličina c_1 i c_2 prema formulama (28) i (30), odnosno (32), može se za velike n približno izvesti pomoću aproksimacije binomne razdiobe $B(n, p_0)$, normalnom razdiobom $N(np_0, np_0(1-p_0))$. U tom slučaju (28) i (30) postaje

(33)
$$\begin{cases} c_1 = np_0 + \Phi^{-1}(\alpha)\sqrt{np_0(1-p_0)}\,, \\ c_2 = np_0 + \Phi^{-1}(1-\alpha)\sqrt{np_0(1-p_0)}\,, \end{cases}$$

dok iz (32) proizlazi

(34)
$$\begin{cases} c_1 = np_0 + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{np_0(1-p_0)} \\ c_2 = np_0 + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{np_0(1-p_0)}\,. \end{cases}$$

Za $n = 50$, $p_0 = 0{,}5$ i $\alpha = 0{,}05$, iz (34) proizlazi, na primjer, da je $c_1 = 18{,}14$ i $c_2 = 31{,}86$, što znači da hipotezu o nultom medijanu prihvaćamo, ako broj članova, koji nisu veći od nule, u danom nizu od 50 podataka, bude između 18 i 32.

5. Wilcoxonov test

Wilcoxonov test uklanja neke nedostatke testa predznaka, ali je primjenljiv samo za testiranje hipoteza o medijanu ($H_0 : M = M_0$), uz pretpostavku da podaci $x_1, \ldots, x_n$ potječu od simetrične kontinuirane razdiobe vjerojatnosti, tj. takve za koju vrijedi

(35)
$$f(M - x) = f(M + x)\,, \qquad x \in \mathbf{R}\,,$$

gdje je f pripadna f.g.v., a M medijan. U terminima f.r.v. F, uvjet simetričnosti razdiobe zapisuje se

(36)
$$F(M - x) = 1 - F(M + x)\,, \qquad x \in \mathbf{R}\,.$$

Bez gubitka općenitosti može se uzeti da je $M_0 = 0$, tako da je riječ o nul-hipotezi $H_0 : M = 0$. Ako je, naime, $M_0 \neq 0$, onda će se na niz podataka $x_1 - M_0, \ldots, x_n - M_0$ primijeniti testiranje hipoteze $H_0 : M = 0$.

Prije nego formalno definiramo Wilcoxonovu statistiku W, opisat ćemo postupak računanja njezine vrijednosti w, na danom konkretnom nizu od $n = 10$ podataka:

$$3{,}2 \quad -1{,}6 \quad 1{,}0 \quad 2{,}1 \quad -0{,}4 \quad 1{,}5 \quad -2{,}6 \quad -0{,}1 \quad 1{,}2 \quad 0{,}5\,.$$

Poredajmo dane brojeve po veličini pripadne apsolutne vrijednosti i načinimo tablicu 3.

Tablica 3.

Apsolutna vrijednost	0,1	0,4	0,5	1,0	1,2	1,5	1,6	2,1	2,6	3,2
Predznak	−	−	+	+	+	+	−	+	−	+
Rang	1	2	3	4	5	6	7	8	9	10

Redni broj u nizu apsolutnih vrijednosti zove se ***rang podataka***. Rangovi pozitivnih podataka su 3,4,5,6,8 i 10, a rangovi negativnih podataka su 1,2,7 i 9. Vrijednost w Wilcoxonove statistike W dobiva se tako da se od zbroja rangova pozitivnih podataka oduzme zbroj rangova negativnih podataka. Dobiva se

$$w = 3 + 4 + 5 + 6 + 8 + 10 - (1 + 2 + 7 + 9) = 17\,.$$

Općenito govoreći, svakom članu niza podataka pripada odgovarajući rang r_i, koji označuje redni broj toga člana pri nizanju apsolutnih vrijednosti od manjih prema većima. Broj r_i može se shvatiti i kao vrijednost diskretne slučajne varijable R_i, koja označuje rang člana X_i u slučajnom uzorku $(X_1, \ldots, X_n)$.

Definirajmo slučajnu varijablu Z_i tako da stavimo

(37)
$$Z_i = \begin{cases} -1\,, & \text{za } X_i \leq 0 \\ 1\,, & \text{za } X_i > 0 \end{cases}\,, \qquad i = 1, \ldots, n\,,$$

i tada se slučajna varijabla

(38)
$$W = \sum_{i=1}^{n} Z_i R_i$$

zove *Wilcoxonova statistika*.

Očigledno je minimalna vrijednost Wilcoxonove statistike

$$w_{\min} = -(1 + 2 + \ldots + n) = -\frac{n(n+1)}{2}\,,$$

a maksimalna vrijednost

$$w_{\max} = 1 + 2 + \ldots + n = \frac{n(n+1)}{2}\,.$$

U maloprije navedenom primjeru ($n = 10$) vrijednost Wilcoxonove statistike iznosi $w = 17$, dok je minimalna vrijednost -55, a maksimalna vrijednost 55, pa se postavlja pitanje da li dobivena vrijednost upućuje na prihvaćanje ili na odbacivanje hipoteze o nultom medijanu.

Ako je hipoteza $H_0 : M = 0$ stvarno istinita, onda se može očekivati podjednaki broj plusova (+) i minusa (−) kod velikih i kod malih, po apsolutnoj vrijednosti, članova niza danih podataka, što će rezultirati vrijednošću Wilcoxonove statistike blizu nule. Dobije li se w blizu w_{min}, onda to znači da predznak minus preteže kod velikih, po apsolutnoj vrijednosti, članova niza podataka, što upućuje na zaključak da podaci potječu iz vjerojatnosne razdiobe koja ima negativni medijan. Ako je pak w blizu w_{max}, onda to znači da predznak plus preteže kod velikih, po apsolutnoj vrijednosti, članova niza podataka i stoga će to upućivati na prihvaćanje alternativne hipoteze $H_1 : M > 0$.

Za određivanje kritičnog područja zadane razine značajnosti α, pri testiranju nul-hipoteze $H_0 : M = 0$, prema alternativnoj hipotezi $H_1 : M < 0$ (ili $H_1 : M > 0$, ili $H_1 : M \neq 0$), nužno je poznavati razdiobu vjerojatnosti Wilcoxonove statistike W, definirane u (38), u uvjetima istinitosti hipoteze H_0.

Uočimo najprije da je W diskretna slučajna varijabla s pripadnim skupom vrijednosti (v. IV.1)

$$A_W = \{-k, -k+2, \dots, k-2, k\}, \qquad k = \frac{n(n+1)}{2},$$

pa ostaje da se odrede pripadne vjerojatnosti

$$P(W = k - 2i), \qquad i = 0, 1, \dots, k.$$

Jedan od načina zasniva se na sljedećim činjenicama:

1. Iz pretpostavke da je promatrana (mjerena) slučajna varijabla kontinuiranog tipa, sa simetričnom oko nule razdiobom vjerojatnosti, proizlazi da je

$$P(X_i \le 0) = P(Z_i = -1) = P(X_i > 0) = P(Z_i = 1) = 0{,}5.$$

2. $Z_1, \dots, Z_n$ nezavisne su slučajne varijable.

3. Slučajne varijable $R_1, \dots, R_n$ nezavisne su u odnosu na $Z_1, \dots, Z_n$.

Iz navedenih činjenica proizlazi da Wilcoxonovoj statistici W pripada ista razdioba vjerojatnosti (v. zad. 13) kao i slučajnoj varijabli

$$V = \sum_{i=1}^{n} V_i, \tag{39}$$

gdje su $V_1, \dots, V_n$ nezavisne slučajne varijable i vrijedi

$$P(V_i = i) = P(V_i = -i) = 0{,}5, \qquad i = 1, \dots, n. \tag{40}$$

Za dani $n \ge 2$, skup $A_W = A_V$ ima 2^n članova, tako da za $n = 2$ imamo

$$A_V = \{-3, -1, 1, 3\},$$

pa neposrednom primjenom (39) i (40) dobivamo

$$P(V = -3) = P(V = -1) = P(V = 1) = P(V = 3) = 0{,}25.$$

Svaka se, naime, vrijednost $(-3, -1, 1, 3)$ slučajne varijable V može, prema (39), realizirati samo na jedan način. Tako je

$$-3 = -1 - 2, \quad -1 = 1 - 2, \quad 1 = -1 + 2, \quad 3 = 1 + 2.$$

Iz (40), pak, proizlazi da svakoj vrijednosti pripada vjerojatnost $\frac{1}{4} = 0{,}25$.

Slično se izrađuju tablice za vrijednosti i pripadne vjerojatnosti slučajne varijable V, odnosno W, za $n = 3, 4, \dots$ (v. tabl. 4).

Tablica 4.

$n = 3$	i	-6	-4	-2	0	2	4	6				
	$P(V = i)$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$				
$n = 4$	i	-10	-8	-6	-4	-2	0	2	4	6	8	10
	$P(V = i)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$

Očigledno je da se povećanjem broja n sve više komplicira (v. zad. 14) izračunavanje vjerojatnosti za slučajnu varijablu V, pa se logično nameće pitanje o približnoj razdiobi vjerojatnosti za velike n. Iz (39) i (40) odmah slijedi da je

$$E[W] = E[V] = 0, \tag{41}$$

$$D[W] = D[V] = \frac{n(n+1)(2n+1)}{6}, \tag{42}$$

pa se, primjenom centralnoga graničnog teorema (v. VI.8) može uzeti da za velike n približno vrijedi

$$Z = W\sqrt{\frac{6}{n(n+1)(2n+1)}} \sim N(0, 1). \tag{43}$$

Pri testiranju hipoteze $H_0 : M = 0$, prema alternativnoj hipotezi $H_1 : M < 0$, kritično područje razine značajnosti α bit će određeno nejednakošću

$$z = w\sqrt{\frac{6}{n(n+1)(2n+1)}} \le \Phi^{-1}(\alpha). \tag{44}$$

Uzme li se $H_1 : M > 0$, kao alternativna hipoteza, kritično je područje određeno nejednakošću

$$w\sqrt{\frac{6}{n(n+1)(2n+1)}} \ge \Phi^{-1}(1 - \alpha), \tag{45}$$

dok je za alternativnu hipotezu $H_1 : M \neq 0$ kritično područje određeno nejednakošću

$$|w|\sqrt{\frac{6}{n(n+1)(2n+1)}} \geq \Phi^{-1}\left(1-\frac{\alpha}{2}\right). \tag{46}$$

Iako se (43), pa stoga i (44), (45) i (46), praktički primijenjuju za $n > 20$, primijenit ćemo ih na podatke iz tabl. 3. Imali smo $n = 10$ i $w = 17$, iz čega se dobiva $z = 0{,}866$. Uzme li se $\alpha = 0{,}05$, dobiva se $\Phi^{-1}(\alpha) = -1{,}65$, $\Phi^{-1}(1-\alpha) = 1{,}65$ i $\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = 1{,}96$, pa se, na temelju (44), (45) i (46), zaključuje da nul-hipotezu treba prihvatiti u odnosu na svaku od tri navedene alternativne hipoteze. Vrijednost $w = 17$ Wilcoxonove test-statistike W ne daje osnovu za odbacivanje hipoteze da je medijan pripadne vjerojatnosne razdiobe nula.

Do istog rezultata došlo bi se i primjenom tabl. IX. iz Dodatka, gdje su prikazane vrijednosti za rubne točke kritičnog područja, izračunane na temelju stvarne razdiobe vjerojatnosti Wilcoxonove statistike W. Tako, na primjer, za $n = 10$ i $\alpha = 0{,}05$ očitavamo $c = 39$, što znači da je $\langle -\infty, -39] \cup [39, \infty\rangle$ kritično područje pri testiranju hipoteze $H_0 : M = 0$, prema alternativnoj hipotezi $H_1 : M \neq 0$. Budući da vrijednost $w = 17$ ne pada u kritično područje, nema razloga da se odbaci H_0.

Primjedba

Već je istaknuto da se Wilcoxonov test primjenjuje samo na kontinuirane slučajne varijable, tako da se u nizu podataka $x_1, \ldots, x_n$ mogu pojaviti jednaki, po apsolutnoj vrijednosti, brojevi samo s vjerojatnošću nula, što praktički znači da se ne očekuje pojava jednakih članova u pripadnom nizu apsolutnih vrijednosti podataka. U praksi se, međutim, ipak mogu pojaviti jednaki članovi (nepreciznost mjerenja i sl.), pa se postavlja pitanje kako pridružiti odgovarajući rang jednakim članovima toga niza. Kako se to radi objasnit ćemo (v. tabl. 5) na nizu podataka

$$0{,}3 \quad -0{,}3 \quad 0{,}2 \quad -0{,}7 \quad 0{,}3 \quad 0{,}7 \quad 0{,}8\,.$$

Rang 3 za članove 0,3 , 0,3 , 0,3 određen je tako da je uzeta aritmetička sredina brojeva 2, 3 i 4. Isto tako je rang 5,5 , za članove 0,7 i 0,7 niza apsolutnih vrijednosti, određen tako da je uzeta aritmetička sredina brojeva 5 i 6.

Tablica 5.

Niz apsolutnih vrijednosti	0,2	0,3	0,3	0,3	0,7	0,7	0,9
Predznak	+	+	+	−	+	−	+
Redni broj	1.	2.	3.	4.	5.	6.	7.
Rang	1	3	3	3	5,5	5,5	7

4. primjer

Da bi se na praktičnu situaciju iz 3. primjera primijenio Wilcoxonov test, potrebno je naprije izračunati razlike (x_i) brojeva iz prvog i drugog retka tabl. 2, čime se dobiva niz podataka

$$19 \quad 13 \quad 8 \quad -15 \quad -37 \quad 29 \quad 18 \quad -18 \quad 14 \quad 26\,,$$

iz čega proizlazi tabl. 6

Tablica 6.

Niz apsolutnih vrijednosti	8	13	14	15	18	18	19	26	29	37
Predznak	+	+	+	−	−	+	+	+	+	−
Rang	1	2	3	4	5,5	5,5	7	8	9	10

Sada možemo izračunati vrijednost Wilcoxonove statistike

$$w = 1+2+3+5{,}5+7+8+9-(4+5{,}5+10) = 16\,.$$

Budući da je riječ o testiranju hipoteze $H_0 : M = 0$, prema alternativnoj hipotezi $H_1 : M \neq 0$, uz razinu značajnosti od 10 %, možemo primijeniti tabl. IX. iz Dodatka ($n = 10, \alpha = 0{,}10$), iz koje se očitava da je pripadno kritično područje $\langle -\infty, -33] \cup [33, \infty\rangle$, pa zaključujemo da hipotezu H_0 ne treba odbaciti.

Kao i primjenom testa predznaka, i primjenom Wilcoxonova testa zaključuje se da ne postoji značajna razlika u kvaliteti auto-guma marke A i marke B.

6. Mann-Whitney-Wilcoxonov test

U IX.6. opisana je primjena hikvadrat-testa pri testiranju hipoteze o jednakosti dvije diskretne razdiobe vjerojatnosti, uz napomenu da se ta metoda može primijeniti i na kontinuirane razdiobe vjerojatnosti, tako da se podaci prethodno grupiraju u razrede. Budući da je grupiranje podataka u razrede, zbog teorijske neutemeljenosti, nepoželjan proces, a sama primjena hikvadrat-testa zahtijeva velik broj podataka, svakako je poželjno imati test za testiranje hipoteze o jednakosti vjerojatnosnih razdioba, koji će zahtijevati slabije pretpostavke.

Neka su, dakle, kao i u IX.6, $x_1, \ldots, x_m$ i $y_1, \ldots, y_n$ dva niza podataka koji potječu od nezavisnih kontinuiranih slučajnih varijabli X i Y, kojima pripadaju funkcije razdiobe vjerojatnosti F i G. Postavlja se zadatak konstruiranja testa za testiranje nul-hipoteze

$$H_0 : F(x) = G(x)\,, \qquad \forall x \in \mathbf{R}\,, \tag{47}$$

prema alternativnoj hipotezi

$$H_1 : F(x) \leq G(x) \text{ ili } F(x) \geq G(x)\,, \qquad \forall x \in \mathbf{R}\,. \tag{48}$$

Primijetimo da se kao alternativna hipoteza pojavljuje iskaz o translaciji ulijevo, odnosno udesno, jedne razdiobe vjerojatnosti s obzirom na drugu.

Test-statistika će se ovdje definirati pomoću rangova podataka u združenom nizu podataka

(z) $$z_1, \ldots, z_{m+n},$$

koji je načinjen tako da su nanizani po veličini članovi nizova

(x) $$x_1, \ldots, x_m$$

i

(y) $$y_1, \ldots, y_n .$$

Kaže se da podatak z_r $(r = 1, \ldots, m+n)$ ima *rang r*.

Pretpostavimo da je $m \leq n$, pa će se zbroj $\widehat{w}$ rangova svih članova niza (x) uzeti kao vrijednost statistike $\widehat{W}$.

5. primjer

Uzmimo da je niz iksova

(x) 55 37 20 41 50 $(m = 5)$,

a niz ipsilona

(y) 85 43 34 65 90 35 75 21 $(n = 8)$.

Združeni niz podataka poredanih po veličini (niz zeova) glasi

(z) 20 21 34 35 37 41 43 50 55 65 75 85 90.

Uoči li se u tom nizu samo pripadnost dotičnog člana nizu (x), odnosno nizu (y), i pripadni rang, dobiva se tabl. 7.

Tablica 7.

Član	x	y	y	y	x	x	y	x	x	y	y	y	y
Rang	1	2	3	4	5	6	7	8	9	10	11	12	13

Vidimo da članovi niza (x) imaju u nizu (z) rangove 1,5,6,8 i 9, dok članovi niza (y) imaju rangove 2,3,4,7,10,11,12 i 13. Pripadna vrijednost test-statistike $\widehat{W}$ jest

$$\widehat{w} = 1 + 5 + 6 + 8 + 9 = 29,$$

pa se odmah postavlja pitanje da li ona indicira odbacivaje ili prihvaćanje hipoteze H_0.

Općenito se najmanja moguća vrijednost $\widehat{w}_{\min}$ test-statistike $\widehat{W}$ dobiva onda kada je svaki član niza (x) manji od svakog člana niza (y) i tada je

$$\widehat{w}_{\min} = 1 + 2 + \ldots + m = \frac{m(m+1)}{2}.$$

Odmah se vidi da najveća moguća vrijednost test-statistike $\widehat{W}$ glasi

$$\widehat{w}_{\max} = n + 1 + n + 2 + \ldots + n + m = mn + \frac{m(m+1)}{2},$$

a postiže se onda kada je svaki član niza (x) veći od svakog člana niza (y).

Test-statistika $\widehat{W}$ očigledno je diskretna slučajna varijabla s pripadnim skupom vrijednosti

$$A_{\widehat{W}} = \{\widehat{w}_{\min}, \widehat{w}_{\min} + 1, \ldots, \widehat{w}_{\max}\},$$

koji u konkretnom primjeru glasi

$$A_{\widehat{W}} = \{15, 16, 17, \ldots, 54, 55\}.$$

U uvjetima istinitosti hipoteze H_0 ne očekujemo ni premalene ni prevelike vrijednosti test-statistike $\widehat{W}$. Da bi se moglo egzaktno odrediti kritično područje zadane razine značajnosti, treba poznavati razdiobu vjerojatnosti slučajne varijable $\widehat{W}$. Za malene m i n mogu se, vrlo lako, izračunati vjerojatnosti $P(\widehat{W} = k)$, $k \in A_{\widehat{W}}$ (v. zad. 17), neposrednim prebrajanjem svih mogućih ishoda i svih povoljnih ishoda za događaj $\{\widehat{W} = k\}$.

Za velike m i n, pak, približno vrijedi da u uvjetima istinitosti hipoteze H_0

(49) $$\widehat{W} \sim N(\mu_{\widehat{W}}, \sigma^2_{\widehat{W}}),$$

gdje je

(50) $$\mu_{\widehat{W}} = \frac{m(m+n+1)}{2}, \qquad \sigma^2_{\widehat{W}} = \frac{mn(m+n+1)}{12}.$$

Ako je hipoteza H_0 stvarno istinita, onda se očekuje dobivanje vrijednosti $\widehat{w}$, test-statistike $\widehat{W}$, u intervalu $\langle c_1, c_2 \rangle$ $(c_1 < c_2)$, za koji vrijedi

(51) $$P(\widehat{W} \leq c_1) = P(\widehat{W} \geq c_2) = \frac{\alpha}{2},$$

tako da će kritično područje razine značajnosti α biti $\langle -\infty, c_1] \cup [c_2, \infty\rangle$.

Veličine c_1 i c_2, za malene m i n, prikazane su u tabl. X. u Dodatku, dok se za velike m i n $(m, n > 10)$ mogu dobiti primjenom (49) i (50).

Uzmu li se podaci iz 5. primjera $(m = 5, n = 8)$, odmah se vidi da je $c_1 = 23$ i $c_2 = 47$, što znači da je $\langle -\infty, 23] \cup [47, \infty\rangle$ kritično područje pri testiranju hipoteze H_0 (podaci potječu od iste razdiobe vjerojatnosti), prema alternativnoj hipotezi H_1 (postoji translacija jedne razdiobe s obzirom na drugu). Budući da dobivena vrijednost $(\widehat{w} = 29)$ test-statistike $\widehat{W}$ ne pada u kritično područje, nema razloga za odbacivanje hipoteze H_0.

7. Medijan-test

Za testiranje hipoteze H_0, da podaci $x_1, \ldots, x_m$ i $y_1, \ldots, y_n$ potječu od slučajnih varijabli X i Y kojima pripada ista razdioba vjerojatnosti, prema alternativnoj hipotezi H_1, da postoji translacija jedne razdiobe s obzirom na drugu, može se iskoristiti i test zasnovan na statistici V, koja ima značenje broja članova niza (x) u lijevoj (donjoj) polovini niza (z) ((x), (y) i (z) definirani su u XIV.6). Ako je $m+n$ paran broj, onda je jasno da lijevu polovinu čini prvih $\frac{m+n}{2}$ članova niza (z), a ako je $m+n$ neparan broj, onda ćemo prvih $\frac{m+n-1}{2}$ članova niza (z) smatrati lijevom polovinom niza (z). Stavimo stoga

$$(52)\qquad k = \begin{cases} \dfrac{m+n}{2} & \text{, za } m+n \text{ parno} \\ \dfrac{m+n-1}{2} & \text{, za } m+n \text{ neparno.} \end{cases}$$

Ako je hipoteza H_0 stvarno istinita, onda se ne očekuje ni premalena ni prevelika vrijednost v statistike V. Očekuje se približno $\frac{m}{2}$ iksova u svakoj polovini niza (z). Kada bi se pojavio premali, ili preveliki, udio iksova u lijevoj polovini niza (z), onda bi se moglo očekivati da oni potječu od razdiobe vjerojatnosti koja je pomaknuta udesno, odnosno ulijevo, s obzirom na razdiobu vjerojatnosti slučajne varijable Y.

Očigledno je V diskretna slučajna varijabla kojoj pripada skup vrijednosti

$$(53)\qquad A_V = \{0, 1, \ldots, \min\{m, k\}\},$$

dok su pripadne vjerojatnosti (v. zad. 19) dane formulom

$$(54)\qquad \mathrm{P}(V = v) = \frac{\binom{m}{v}\binom{n}{k-v}}{\binom{m+n}{k}}, \qquad v \in A_V .$$

Formulama (53) i (54) definirana je tzv. *hipergeometrijska razdioba* s parametrima m, n i k ($m, n, k \in \mathbf{N}$, $k < m+n$) (v. [34]).

Test utemeljen na statistici V zove se *medijan-test*.

Formula (54) omogućuje da se, za danu razinu značajnosti α, bar približno odrede brojevi c_1 i c_2, tako da vrijedi

$$(55)\qquad \mathrm{P}(V \le c_1) = \mathrm{P}(V \ge c_2) = \frac{\alpha}{2} .$$

Poznata je činjenica (v. [11]) da se za velike m i n hipergeometrijska razdioba može aproksimirati normalnom razdiobom $\mathrm{N}(\mu_V, \sigma_V^2)$, gdje je

$$(56)\qquad \mu_V = \mathrm{E}[V] = \frac{km}{m+n}, \qquad \sigma_V^2 = \mathrm{D}[V] = \frac{kmn(m+n-k)}{(m+n)^2(m+n-1)},$$

što bitno olakšava određivanje rubova c_1 i c_2 kritičnog područja, prema zahtjevima (55).

Formule (56) postaju još jednostavnije kada je $k = \frac{m+n}{2}$. Tada je

$$(56a)\qquad \mathrm{E}[V] = \frac{m}{2}, \qquad \mathrm{D}[V] = \frac{mn}{4(m+n-1)} .$$

Dobije li se, na danim podacima, vrijednost v, test-statistike V iz intervala $\langle c_1, c_2 \rangle$, hipoteza H_0 će se prihvatiti, dok će se u protivnom odbaciti.

Ako se sumnja samo na mogućnost pomaka ulijevo razdiobe vjerojatnosti slučajne varijable X, s obzirom na razdiobu vjerojatnosti slučajne varijable Y, onda će se nul-hipotezi $H_0 : F(x) = G(x)$, kao alternativna postaviti hipoteza $H_1 : F(x) \ge G(x)$ i tada će kritično područje biti interval $[c, \infty\rangle$, gdje je c određeno uvjetom

$$(57)\qquad \mathrm{P}(V \ge c) = \alpha .$$

Kada se dobije previše iksova (više od c) u donjoj polovini združenoga uređenog niza podataka, hipoteza H_0 će se odbaciti i prihvatiti hipoteza H_1.

6. primjer

Mjerenjem slučajne varijable X dobiven je niz od $m = 9$ podataka

(x) 45 50 52 44 57 53 55 53 54 ,

a mjerenjem slučajne varijable Y niz od $n = 11$ podataka

(y) 59 39 47 59 44 62 61 62 64 54 49 .

Može li se smatrati da oba niza potječu od iste razdiobe vjerojatnosti?

Zadatak se, naravno, može formulirati kao testiranje hipoteze H_0 iz (47), uz razinu značajnosti, recimo, $\alpha = 0{,}05$. U tu svrhu formirajmo združeni niz (z), po veličini poredanih podataka

(z)

39	44	44	45	47	49	50	52	53	53
y	x	y	x	y	y	x	x	x	x
54	54	55	57	59	61	62	62	62	64
x	y	x	x	y	y	y	y	y	y .

U danom je primjeru $k = \frac{m+n}{2} = 10$, pa odmah vidimo da među prvih deset članova niza (z) ima $v = 6$ članova niza (x), a očekuje ih se $\frac{m}{2} = 4{,}5$.

Da bi se odredilo kritično područje zadane razine značajnosti na temelju test-statistike V, uzet ćemo da približno vrijedi $V \sim \mathrm{N}(\mu_V, \sigma_V^2)$, gdje je, prema (56a), $\mu_V = 4{,}5$ i $\sigma_V^2 = 1{,}30$. Iz (55) se tada dobiva

$$c_1 = \sigma_V\, \Phi^{-1}\Big(\frac{\alpha}{2}\Big) + \mu_V \approx 1{,}14(-1,96) + 4{,}5 = 2{,}27 ,$$

$$c_2 = \sigma_V\, \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) + \mu_V \approx 1{,}14 \cdot 1,96 + 4{,}5 = 6{,}73 ,$$

tako da je kritično područje $\langle -\infty; 2{,}27] \cup [6{,}73; \infty\rangle$, pa vidimo da vrijednost ($v = 6$) test-statistike V ne pada u kritično područje, što upućuje na prihvaćanje hipoteze o jednakosti vjerojatnosnih razdioba slučajnih varijabli X i Y.

Zanimljivo je primijetiti da bismo za testiranje hipoteze H_0 mogli primijeniti i hikvadrat-test, onako kako je to opisano u IX.6. Mogli bismo, naime, smatrati da prvih deset članova niza (z) pripada prvom, a preostalih deset članova drugom razredu, pri grupiranju podataka u razrede ($r = 2$).

Frekvencija iksova u prvom razredu je $\widehat{f}_1 = 6$, a u drugom $\widehat{f}_2 = 3$, dok su odgovarajuće frekvencije ipsilona $\widehat{g}_1 = 4$ i $\widehat{g}_2 = 7$. Izračuna li se vrijednost d test-statistike D (formula (42) u IX.6), dobiva se

$$d = \frac{1}{mn}\left[\frac{(n\widehat{f}_1 - m\widehat{g}_1)^2}{\widehat{f}_1 + \widehat{g}_1} + \frac{(n\widehat{f}_2 - m\widehat{g}_2)^2}{\widehat{f}_2 + \widehat{g}_2}\right] = $$

$$= \frac{1}{99 \cdot 10}\left[(11 \cdot 6 - 9 \cdot 4)^2 + (11 \cdot 3 - 9 \cdot 7)^2\right] \approx 1{,}81\,.$$

U IX.6. navedeno je da statistici D, uz uvjet istinitosti hipoteze H_0, pripada hikvadrat-razdioba sa $r - 1$ stupnjeva slobode, što u ovom slučaju ($r = 2$) znači da je kritično područje određeno nejednakošću

$$d \geq H_1^{-1}(1 - \alpha) = H_1^{-1}(0{,}95) = 3{,}84\,.$$

Odmah se vidi da dobivena vrijednost (1,81) test-statistike D ne pada u kritično područje $[3{,}84; \infty\rangle$, pa se i ovim testom sugerira zaključak o jednakosti promatranih razdioba vjerojatnosti.

Određenom prilagodbom opisani se test može iskoristiti i za testiranje hipoteze o stacionarnosti niza podataka, prema alternativnoj hipotezi da postoji određeni trend pri nizanju podataka $x_1, \ldots, x_n$. Pretpostavimo da je $n = 2m$ ($m \in \mathbf{N}$), pa se tada dani niz podataka može rastaviti na dva podniza

(L) $$x_1, \ldots, x_m$$

(D) $$x_{m+1}, \ldots, x_n\,.$$

Neka slučajna varijabla V označuje broj onih članova podniza (L) koji su manji od vrijednosti uzoračkog medijana $\widehat{m}$ (v. (11)), pa se odmah vidi da je V diskretna slučajna varijabla kojoj pripada skup vrijednosti

(58) $$A_V = \{0, 1, \ldots, m\}\,,$$

a u uvjetima istinitosti nul-hipoteze H_0 :(mjerenja potječu od nezavisnih slučajnih varijabli iste razdiobe vjerojatnosti) slučajnoj varijabli V pripada diskretna razdioba vjerojatnosti zadana formulom

(59) $$P(V = v) = \frac{\binom{m}{v}\binom{m}{m-v}}{\binom{2m}{m}}\,, \qquad v \in A_V\,.$$

Usporedbom (58) i (59) sa (53) i (54) vidi se da su formule (58) i (59) posebni slučaj formula (53) i (54), tj. da se radi o hipergeometrijskoj razdiobi, pa iz (56a) proizlazi da je u ovom slučaju

(60) $$\mathrm{E}[V] = \frac{m}{2}\,, \qquad \mathrm{D}[V] = \frac{m^2}{4(2m-1)}\,.$$

Odrede li se rubovi c_1 i $c_2 (c_1 < c_2)$ kritičnog područja razine značajnosti α na temelju (55) i (59), odnosno na temelju aproksimacije hipergeometrijske razdiobe odgovarajućom normalnom razdiobom, hipoteza H_0 prihvatiti će se onda kada vrijednost v, test-statistike V, padane u interval $\langle c_1, c_2\rangle$. U protivnom smatrat će se da postoji određeni trend (rastući ili padajući) u danom nizu statističkih podataka.

Medijan-test, prilagođen testiranju hipoteze o postojanju, odnosno nepostojanju, određenog trenda u danom nizu statističkih podataka zove se *Mood-Brownov medijan-test.*

7. primjer

Bilježeći maksimalni godišnji riječni vodostaj u razdoblju od $n = 10$ godina, dobiven je ovaj niz statističkih podataka:

(L)	346	306	448	402	439
	4.	3.	10.	6.	9.
(D)	390	418	150	231	437 .
	5.	7.	1.	2.	8.

Ispod svake vrijednosti vodostaja naveden je odgovarajući redni broj u rastućem nizu tih podataka.

Smije li se taj niz smatrati stacionarnim nizom u smislu da ne postoji određeni trend? Odgovor ćemo potražiti primjenom Mood-Brownova medijan-testa.

U danom je primjeru $n = 10$ ($m = 5$), dok je vrijednost uzoračkog medijana $\widehat{m} = 402$. Prema (11) to je šesta po veličini vrijednost u danom nizu podataka.

Odmah se vidi da je vrijednost test-statistike V u ovom primjeru $v = 2$, jer u podnizu (L) ima dva člana manja od $\widehat{m} = 402$.

Izračunaju li se, primjenom (59), odgovarajuće vjerojatnosti za slučajnu varijablu V, dobiva se

$$P(V = 0) = P(V = 5) = \frac{1}{252} \approx 0{,}004$$

$$P(V = 1) = P(V = 4) = \frac{25}{252} \approx 0{,}100$$

$$P(V = 2) = P(V = 3) = \frac{100}{252} \approx 0{,}396\,.$$

Odmah se vidi da je $P(V \leq 1) = P(V \geq 4) \approx 0{,}104$, tako da kritičnom području $\langle -\infty, 1] \cup [4, \infty\rangle$ pripada razina značajnosti $\alpha = 0{,}208$. Dobivena vrijednost ($v = 2$) test-statistike V ne pada ni u to kritično područje, što upućuje na prihvaćanje hipoteze o nepostojanju trenda u danom nizu statističkih podataka.

Da smo se poslužili aproksimacijom hipergeometrijske razdiobe normalnom razdiobom $\mathrm{N}\left(\frac{5}{2}, \frac{25}{36}\right)$, što baš nije preporučljivo zbog premalenog n ($n = 10$), za rubove kritičnog područja razine značajnosti $\alpha = 0{,}10$ dobili bismo vrijednosti

$$c_1 = \sigma\,\Phi^{-1}\left(\frac{\alpha}{2}\right) + \mu = \frac{5}{6}\cdot(-1{,}65) + \frac{5}{2} = 1{,}125\,,$$

$$c_2 = \sigma\,\Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \mu = \frac{5}{6}\cdot 1{,}65 + \frac{5}{2} = 3{,}875\,,$$

pa vidimo da i uz takav račun vrijednost test-statistike ne pada u kritično područje.

8. Test-serija

Za testiranje hipoteze H_0 o jednakosti dviju razdioba vjerojatnosti, na temelju dvaju nizova statističkih podataka $x_1, \ldots, x_m$ i $y_1, \ldots, y_n$, može se konstruirati test koji se zasniva na intuitivnoj ideji da u uvjetima istinitosti hipoteze H_0 iksovi i ipsiloni trebaju biti "izmiješani na slučajan način" u združenom nizu, po veličini poredanih podataka. Kao pokazatelj izmiješanosti iksova i ipsilona u združenom nizu (v. tabl. 7) može poslužiti broj serija u tom nizu. Tako u nizu u tabl. 7. imamo najprije seriju od jednog iksa, zatim seriju od tri ipsilona, pa seriju od dva iksa itd. Vidimo da ima ukupno 6 serija u tom nizu.

Najmanji je mogući broj serija 2, koji se s podacima iz 5. primjera postiže u ovim situacijama:

(a) x x x x x y y y y y y y y

(b) y y y y y y y y x x x x x .

Očigledno je da obje situacije (a) i (b) upućuju na odbacivanje hipoteze H_0, jer dani podaci pokazuju tendenciju pomaka razdiobe iksova s obzirom na razdiobu ipsilona.

Uzme li se kao test-statistika broj R serija (engleski: runs) u združenom nizu (z), po veličini poredanih podataka iz niza (x) i niza (y), hipoteza H_0 će se odbaciti onda kada se dobije premalena vrijednost test-statistike R.

Da bi se odredilo kritično područje zadane razine značajnosti α, treba poznavati razdiobu vjerojatnosti diskretne slučajne varijable R. Pokazuje se (v. zad. 24) da u uvjetima istinitosti hipoteze H_0 vrijedi

$$\mathrm{P}(R = r) = \begin{cases} \dfrac{1}{\binom{m+n}{m}}\left[\binom{m-1}{k}\binom{n-1}{k-1} + \binom{m-1}{k-1}\binom{n-1}{k}\right], & r = 2k+1 \\[2ex] \dfrac{2}{\binom{m+n}{m}}\binom{m-1}{k-1}\binom{n-1}{k-1}, & r = 2k\,, \end{cases} \tag{61}$$

gdje je $k \in \{1, 2, \ldots, \min\{m, n\}\}$ i r označava broj serija iksova (ipsilona) u združenom nizu (z). Za $m, n > 10$ približno vrijedi da $R \sim \mathrm{N}(\mu_R, \sigma_R^2)$, gdje je

$$\mu_R = \frac{2mn}{m+n} + 1\,, \qquad \sigma_R^2 = \frac{(\mu_R - 1)(\mu_R - 2)}{m+n-1}\,, \tag{62}$$

što omogućuje jednostavno određivanje veličine c, za koju vrijedi $\mathrm{P}(R \le c) = \alpha$. Očigledno je

$$c = \sigma_R\,\Phi^{-1}(\alpha) + \mu_R\,, \tag{63}$$

tako da će se hipoteza o jednakosti promatranih razdioba vjerojatnosti odbaciti kada vrijednost r, test-statistike R, padne u kritično područje $\langle -\infty, c]$.

Za malene vrijednosti m i n ($m, n < 10$) izrađene su tablice (v. tabl. XI. u Dodatku) koje omogućuju određivanje rubova kritičnog područja.

Primijeni li se test serija na podatke iz 6. primjera, vidimo da je $r = 9$ odgovarajuća vrijednost test-statistike R. Budući da je $m = 9$ i $n = 11$, formule (62) daju $\mu_R = 10{,}9$ i $\sigma_R = 2{,}15$. Uzme li se razina značajnosti $\alpha = 0{,}05$, iz (63) proizlazi da je $c = 7{,}35$, pa vidimo da vrijednost ($r = 9$) test-statistike R ne pada u kritično područje $\langle -\infty; 7{,}35]$.

Prema tome i test-serija, kao i medijan-test, upućuje na zaključak da podaci niza (x) i niza (y) iz 6. primjera potječu od iste razdiobe vjerojatnosti.

Vidjeli smo da slučajna varijabla R općenito pokazuje broj serija u nizu sastavljenom od dva simbola x i y, pa se kao takva može iskoristiti i za testiranje hipoteze o "slučajnosti" (engleski: randomness).

Pogledamo li situacije (a) i (b), gdje je vrijednost slučajne varijable R jednaka 2, odmah zaključujemo da tu nikako ne bismo trebali smatrati da se x i y pojavljuju na slučajan način, jer je očigledno da iza x redovito slijedi x, a iza y gotovo uvijek slijedi y. Pogleda li se, pak, situacija

(c) x y y x y y x y y x y y x,

u kojoj imamo 9 serija, može se posumnjati u slučajnost nizanja iksova i ipsilona, jer se iz (c) naslućuje postojanje zakonitosti (periodičnosti) da iza svakog iksa dolaze dva ipsilona.

Prema tome, hipoteza H_0 o slučajnosti niza satavljenog od m simbola jedne vrste (iksova) i n simbola druge vrste (ipsilona) odbacit će se onda kada se dobije premalena ili prevelika vrijednost test-statistike R. U uvjetima istinitosti hipoteze H_0 diskretnoj slučanoj varijabli R pripada razdioba vjerojatnosti opisana formulama (61), koja se za velike m i n ($m, n > 10$) aproksimira normalnom razdiobom $\mathrm{N}(\mu_R, \sigma_R^2)$, gdje su μ_R i σ_R^2 izraženi formulama (62). To omogućuje da se odredi kritično područje zadane razine značajnosti α za hipotezu H_0.

8. primjer

Generatorom pseudoslučajnih brojeva dobiven je niz

2 1 8 5 2 5 1 5 3 5 8 2 5 0 3 3 5 9 2 3 .

Možemo li ga smatrati slučajnim nizom?

Zamijenimo svaki neparni broj u tom nizu simbolom x, a svaki parni broj simbolom y, pa dobivamo niz

$$\mathrm{y\ x\ y\ x\ y\ x\ x\ x\ x\ x\ y\ y\ x\ y\ x\ x\ x\ x\ y\ x},$$

na kojem možemo registrirati broj serija $r = 12$.

Uzme li se razina značajnosti $\alpha = 0{,}10$, rubovi c_1 i c_2 kritičnog područja odredit će se iz uvjeta

$$\mathrm{P}(R \le c_1) = \mathrm{P}(R \ge c_2) = \frac{\alpha}{2} = 0{,}05\,. \tag{64}$$

Budući da je $m = n = 10$, prema (62) se dobiva $\mu_R = 11$ i $\sigma_R \approx 2{,}18$, tako da približno vrijedi

$$c_1 = \sigma_R \Phi^{-1}\Big(\frac{\alpha}{2}\Big) + \mu_R = 7{,}4\,,$$

$$c_2 = \sigma_R \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) + \mu_R = 14{,}6\,,$$

što znači da je pripadno kritično područje $\langle -\infty; 7{,}4] \cup [14{,}6; \infty\rangle$. Vrijednost ($r = 12$) test-statistike R ne pada u kritično područje, pa zaključujemo da generirani niz od 20 pseudoslučajnih brojeva zadovoljava test serija i u tom smislu ga možemo smatrati slučajnim nizom.

Za potpuniju provjeru slučajnosti nekog niza pseudoslučajnih brojeva obično se primjenjuje više različitih testova (v. [48]). Vidjeli smo da se testom serija uglavnom provjerava postojanje izvjesne periodičnosti, dok se Mood-Brownovim medijan-testom može provjeriti postojanje određenog trenda. Primjenom hikvadrat-testa, kako je opisano u IX.3, IX.6 i IX.7, može se provjeriti hipoteza o uniformnoj razdiobi znamenaka, odnosno hipoteza o homogenosti niza pseudoslučajnih brojeva. Tek nakon "prolaza" na tim, eventualno još i na nekim drugim testovima, smatrat će se da promatrani niz zadovoljava uvjet slučajnosti.

9. Robusne metode

Pojam *robusnosti* (engleski: *robust*) relativno je noviji pojam u matematičkoj statistici (v. [20]), a povezan je i s problemima procjene parametara i s testiranjem statističkih hipoteza. Tako se govori o *robusnim procjeniteljima* i o *robusnim testovima*.

Poznato je da je uzoračka aritmetička sredina $\overline{X}$, kao ML-procjenitelj za parametar $\mu = \mathrm{E}[X]$ normalne razdiobe $\mathrm{N}(\mu, \sigma^2)$, nepristran, konzistentan i najefikasniji procjenitelj. Međutim, uzoračka aritmetička sredina nije najbolji procjenitelj za očekivanje $\mathrm{E}[X] = \beta$ u parametarskom modelu, gdje se kao klasa dopuštenih razdioba uzimaju sve Laplaceove razdiobe s parametrima α i β. U VI.5. je pokazano da je uzorački medijan $\widehat{M}$ ML-procjenitelj za parametar β i kao takav je efikasniji od $\overline{X}$. Uzme li se, pak, jednoparametarska klasa Cauchyjevih razdioba, kojima pripada funkcija gustoće vjerojatnosti

$$f_t(x) = \frac{1}{\pi[1 + (x - t)^2]}\,, \qquad x \in \mathbf{R}\,, \tag{65}$$

kao klasa dopuštenih razdioba, i $\overline{X}$ kao procjenitelj za nepoznati parametar $t \in \mathbf{R}$, odmah se vidi da je uzoračka aritmetička sredina $\overline{X}$ vrlo loš procjenitelj za t. Poznata je, naime, činjenica da Cauchyjeva razdioba nema ni konačno očekivanje ni konačnu varijancu pa se ne može govoriti ni o nepristranosti ni o efikasnosti uzoračke aritmetičke sredine kao procjenitelja za nepoznati parametar t Cauchyjeve razdiobe (65).

Parametar t Cauchyjeve razdiobe, kao i parametri β za Laplaceovu i μ za normalnu razdiobu, parametri su lokacije i sva tri imaju značenje medijana odgovarajuće razdiobe, pa se intuitivno teško prihvaća činjenica da je u prvom slučaju uzoračka aritmetička sredina najbolji procjenitelj za medijan normalne razdiobe, u drugom slučaju može poslužiti kao procjenitelj za medijan Laplaceove razdiobe, ali ima i boljih, dok u trećem slučaju gotovo da i nema smisla uzimati $\overline{X}$ kao procjenitelj za medijan Cauchyjeve razdiobe.

To se zbiva zbog toga što se u normalnoj razdiobi, među članovima niza $x_1, \ldots, x_n$ uzoračkih vrijednosti, ne pojavljuju ***stršeće vrijednosti*** (engleski: *outliers*), jer je poznato da je praktički nevjerojatno da se izvan intrevala $\langle \mu - 3\sigma, \mu + 3\sigma\rangle$ izmjeri vrijednost slučajne varijable $X \sim \mathrm{N}(\mu, \sigma^2)$. Kaže se da normalna razdioba ima ***kratke repove***, za razliku od Laplaceove, a pogotovo Cauchyjeve razdiobe, koje imaju ***duge repove***.

Pojava stršećih vrijednosti u nizu statističkih podataka $x_1, \ldots, x_n$ bitno utječe na vrijednost $\overline{x}$ aritmetičke sredine. Statistika $\overline{X}$, u modelima gdje klasu dopuštenih razdioba čine vjerojatnosne razdiobe dugog repa, ima vrlo veliku varijancu (može biti i ∞), što znači da će kao procjenitelj imati slabu efikasnost. Stoga se prirodno nameće zadatak da se pronađe procjenitelj koji možda neće biti najbolji (najefikasniji procjenitelj) ni za jednu razdiobu vjerojatnosti iz vrlo široke klase (recimo sve dvoparametarske kontinuirane simetrične razdiobe), ali će biti vrlo dobar za sve razdiobe te klase. Takav procjenitelj, ako postoji, zove se robusni procjenitelj.

9. primjer

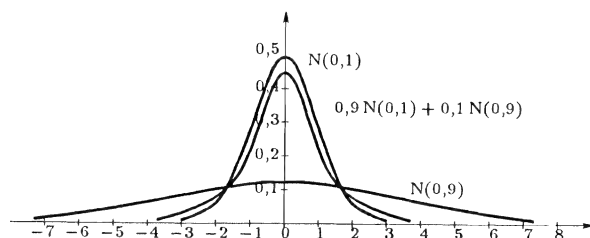
Neka je f_0 f.g.v. za normalnu razdiobu $\mathrm{N}(\mu_0, \sigma^2)$ i f_c f.g.v. za $\mathrm{N}(\mu_0, c\sigma^2)$ ($c > 1$). Neka je nadalje $0 < p < 1$ i $q = 1 - p$, pa se tada razdioba vjerojatnosti sa f.g.v. f, gdje je

$$f(x) = pf_0(x) + qf_c(x)\,, \qquad x \in \mathbf{R}\,, \tag{66}$$

zove *pomućena* (engleski: *contaminated*) *normalna razdioba*.

Ako niz podataka $x_1, \ldots, x_n$ potječe od pomućene normalne razdiobe, onda možemo smatrati da pri svakom mjerenju postoji mogućnost da se dobije vrijednost slučajne varijable $X_0 \sim \mathrm{N}(\mu_0, \sigma^2)$ s vjerojatnošću p, ili vrijednost slučajne varijable $X_c \sim \mathrm{N}(\mu_0, c\sigma^2)$ s vjerojatnošću q. To znači da se u danom nizu podataka može očekivati njih $100p$ % iz $\mathrm{N}(\mu_0, \sigma^2)$ i njih $100q$ % iz $\mathrm{N}(\mu_0, c\sigma^2)$.

Ako je p blizu 1, odnosno q blizu 0, onda se pomućena normalna razdioba "malo razlikuje" od normalne razdiobe $\mathrm{N}(\mu_0, \sigma^2)$, što će se na podacima $x_1, \ldots, x_n$ odraziti tako da će se među njima eventualno pojavljivati stršeće vrijednosti. Glavnina podataka podvrgava se $\mathrm{N}(\mu_0, \sigma^2)$, dok se stršeće vrijednosti podvrgavaju $\mathrm{N}(\mu_0, c\sigma^2)$.



Slika 44. Skica pomućene normalne razdiobe i odgovarajućih normalnih razdioba za $\mu_0 = 0$, $\sigma = 1$, $c = 9$ i $p = 0{,}9$

Označimo li sa X slučajnu varijablu kojoj pripada f.g.v. (66), tada je

$$E[X] = \mu_0 , \qquad V[X] = \sigma^2(p + cq) . \tag{67}$$

Promotri li se pripadna uzoračka aritmetička sredina $\overline{X}$, odmah se vidi da je

$$E[\overline{X}] = \mu_0 , \qquad V[\overline{X}] = \frac{\sigma^2}{n}(p + cq) , \tag{68}$$

iz čega se razabire da je $\overline{X}$ nepristrani procjenitelj za nepoznati parametar μ_0, čija efikasnost (varijanca $V[\overline{X}]$) bitno ovisi o c. Kada je $p = 0{,}9$, $q = 0{,}1$ i $c = 9$ (v. sl. 44), ispada $V[\overline{X}] = 1{,}8\,\frac{\sigma^2}{n}$, pa se vidi da je ta efikasnost tek oko 55 % od efikasnosti aritmetičke sredine kao procjenitelja u klasi nepomućenih normalnih razdioba ($c = 1$).

Uzme li se uzorački medijan $\widehat{M}$ kao procjenitelj za parametar μ_0 u klasi pomućnih normalnih razdioba, onda se njegova efikasnost može približno izraziti (v. (12)) formulom

$$E[(\widehat{M} - \mu_0)^2] = \frac{1}{n}\,\frac{1}{4[f(\mu_0)^2]} = \frac{\pi\sigma^2}{2n}\,\frac{1}{\left(p + \frac{q}{c}\right)^2} , \tag{69}$$

pa se vidi da njegova efikasnost mnogo blaže ovisi o c, nego efikasnost procjenitelja $\overline{X}$. Dapače, za $c \to \infty$, ona ne teži beskonačnosti kao u (68), već teži prema $\frac{\pi\sigma^2}{2np^2}$. S druge strane, za $c = 1$, tj. kada je riječ o nepomućenoj normalnoj razdiobi, efikasnost uzoračkog medijana $\widehat{M}$, kao procjenitelja za parametar μ_0, jest $\frac{\pi\sigma^2}{2n}$, iz čega se vidi da ona iznosi oko 64 % od efikasnosti aritmetičke sredine $\overline{X}$.

Mogli bismo reći da je uzorački medijan prilično robustan procjenitelj za parametar lokacije (medijan, očekivanje) u vrlo širokoj klasi kontinuiranih simetričnih razdioba vjerojatnosti. U klasi normalnih razdioba njegova je efikasnost oko 64 % od efikasnosti uzoračke aritmetičke sredine, koja je u tom slučaju najefikasniji procjenitelj.

Još bolji, u smislu robusnosti, pokazao se procjenitelj definiran pomoću uzoračkih kvartila $\widehat{X}_{0,25}$ i $\widehat{X}_{0,75}$ (v. XIV.2), tako da se stavi

$$\widehat{M}_1 = \frac{1}{2}\,(\widehat{X}_{0,25} + \widehat{X}_{0,75}) . \tag{70}$$

Iz (70) je vidljivo da vrijednost $\widehat{m}_1$ statistike $\widehat{M}_1$ neće biti osjetljiva na stršeće vrijednosti, dok s druge strane procjenitelj $\widehat{M}_1$ zadržava vrlo visoku efikasnost i onda kada podaci potječu od normalne razdiobe (više od 80 % efikasnosti aritmetičke sredine).

Postoje određene metode (v. [20]) kojima se mogu određivati robusni procjenitelji na temelju određenoga općeg načela, poput metode najveće vjerojatnosti. U tom svjetlu robusni se procjenitelji mogu tretirati kao određena generalizacija ML-procjenitelja. Detaljnije obrazloženje postupka određivanja robusnih procjenitelja prelazi okvire ove knjige.

Primijetimo, ipak, da se postupkom *potkresivanja* (engleski: trim, truncate), tj. odbacivanjem određenog postotka (3 % ili 6 % i sl.) najvećih i najmanjih vrijednosti u danom nizu podataka i zatim izračunavanjem aritmetičke sredine i varijance na *potkresanom uzorku*, dobivaju robusne procjene za očekivanje (medijan) i varijancu kao nepoznate parametre. Ti su procjenitelji efikasni u modelu s vrlo opsežnom klasom dopuštenih razdioba, a i u samoj klasi normalnih razdioba zadržavaju vrlo visoku efikasnost.

Da bi se bar donekle razjasnila ideja robusnog testa razmotrit će se idući primjer.

10. primjer

U XIV.4. opisan je problem testiranja nul-hipoteze da spareni podaci (uređeni parovi) (u_i, v_i) $(i = 1, \ldots, n)$ potječu od nezavisnih kontinuiranih slučajnih varijabli U_i i V_i, kojima pripada ista razdioba vjerojatnosti, prema alternativnoj hipotezi da ne potječu od iste razdiobe vjerojatnosti. Ondje je problem rješavan primjenom testa predznaka, dok je u XIV.5. pokazano da se isti problem može rješavati i pomoću Wilcoxonova testa.

Budući da je u oba slučaja nul-hipoteza formulirana kao $H_0 : M = 0$, gdje M označuje medijan razdiobe iz koje potječu podaci $x_i = u_i - v_i$ $(i = 1, \ldots, n)$, nameće se ideja da se na isti problem primijeni test-statistika

$$T = \frac{\overline{X}}{S}\sqrt{n} , \tag{71}$$

gdje je $\overline{X}$ uzoračka aritmetička sredina, a S^2 uzoračka korigirana varijanca za podatke $x_1, \ldots, x_n$. Statistika (71) je, zapravo, posebni slučaj statistike T iz 6. primjera u VIII.6. (formula (70) uz $\mu_0 = 0$), kojoj uz pretpostavku da podaci potječu od normalne razdiobe $N(0, \sigma^2)$ pripada Studentova razdioba sa $n - 1$ stupnjeva slobode. Ako pretpostavka o normalnosti razdiobe nije ispunjena, onda se ne može jednostavno odrediti kritično područje zadane razine značajnosti za test s test-statistikom (71). Nasuprot tome, test predznaka i Wilcoxonov test primjenljivi su na mnogo širu klasu razdioba vjerojatnosti, pa se prirodno nameće pitanje o

njihovoj "efikasnosti" na klasi normalnih razdioba. No, nezgoda je što ovdje nije definiran pojam efikasnosti testa. U IX.4. uveden je, međutim, pojam razlučivosti, koji se može upotrijebiti u navedenu svrhu.

Uzme li se realni broj δ, može se postaviti zadatak da se odredi najmanja veličina $n_0(\delta)$ slučajnog uzorka, kojim se može, uz zadanu razinu značajnosti α, primjenom test-statistike T iz (71), razlučiti normalna razdioba $N(\delta, \sigma^2)$ od normalne razdiobe $N(0, \sigma^2)$. Isto se tako može postaviti zadatak da se odredi najmanja veličina $n_1(\delta)$ slučajnog uzorka kojim se može, uz istu razinu značajnosti α, primjenom testa predznaka, razlučiti $N(\delta, \sigma^2)$ od $N(0, \sigma^2)$. Analogno značenje imat će veličina $n_2(\delta)$ za Wilcoxonov test.

Može se pokazati da vrijedi (v. [4])

$$\lim_{\delta \to 0} \frac{n_0(\delta)}{n_1(\delta)} = \frac{2}{\pi} \approx 0{,}64\,, \tag{72}$$

$$\lim_{\delta \to 0} \frac{n_0(\delta)}{n_2(\delta)} = \frac{3}{\pi} \approx 0{,}95\,, \tag{73}$$

što se može interpretirati tako da se kaže da test predznaka ima 64 % efikasnosti Studentova testa, a Wilcoxonov test da ima 95 % efikasnosti s obzirom na Studentov test, pri testiranju hipoteze o nultom medijanu u modelu s klasom normalnih razdioba $N(\delta, \sigma^2)$ kao klasom dopuštenih razdioba vjerojatnosti.

Pojednostavnjeno govoreći, mogli bismo reći da test predznaka zahtijeva $\frac{\pi}{2} \approx$ $\approx 1{,}57$ puta veći uzorak od Studentova testa, dok Wilcoxonov test zahtijeva samo $\frac{\pi}{3} \approx 1{,}05$ puta veći uzorak od Studentova testa, za isti stupanj razlučivosti na klasi normalnih razdioba. To pokazuje da nije preporučljivo koristiti se testom predznaka u situaciji kada se može smatrati da podaci potječu od normalne razdiobe, jer mu je razlučivost mnogo slabija od razlučivosti Studentova testa. Ostaje, međutim, prednost testa predznaka da nije osjetljiv na odstupanje od normalne razdiobe i da je vrlo jednostavan za primjenu, što nije slučaj sa Studentovim testom.

Wilcoxonov, pak, test ima razlučivost gotovo jednaku razlučivosti Studentova testa na podacima iz normalne razdiobe, a posjeduje i svojstvo neosjetljivosti na odstupanje od normalne razdiobe, tako da je primjenljiv na mnogo širu klasu razdioba vjerojatnosti i u tom smislu se smatra ***robusnim testom***.

Zadaci

1. Izvedite formulu (6) i (7).
Uputa: Primijenite približnu formulu za $E[\widehat{\Sigma}]$ iz VI.6. i činjenicu da su $\overline{X}$ i $\widehat{\Sigma}$ nezavisne slučajne varijable.

2. Nađite ML-procjenitelj za kvantil p-tog reda, odredite njegovo očekivanje i očekivanu kvadratnu grešku, uz pretpostavku da podaci potječu iz uniformne razdiobe $U(0, a)$ $(a > 0)$.

3. Odgovorite na pitanje iz zad. 2, uz pretpostavku da podaci potječu iz eksponencijalne razdiobe $Ex(\alpha)$ $(\alpha > 0)$.

4. Usporedite očekivanu kvadratnu grešku za procjenitelja iz parametarskog modela (formula (3)) i procjenitelja za medijan iz neparametarskog modela (formula (9)), uz pretpostavku da je n (veličina uzorka) veliko i da podaci potječu iz:

a) normalne razdiobe,
b) uniformne razdiobe,
c) eksponencijalne razdiobe.

5. U kakvom su odnosu očekivana kvadratna greška pri procjeni kvantila p-tog reda eksponencijalne razdiobe $Ex(\alpha)$ za $p = 0{,}01$ i $p = 0{,}99$.

6. Odredite interval povjerenja zadane pouzdanosti γ za kvantil x_p, uz primjenu asimptotske normalnosti procjenitelja $\widehat{X}_p$, u parametarskom modelu s klasom dopuštenih razdioba:

a) $\mathcal{P} = \{U(0, a) : a > 0\}$,
b) $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbf{R},\ \sigma > 0\}$,
c) $\mathcal{P} = \{Ex(\alpha) : \alpha > 0\}$.

Uputa: Primijenite postupke i formule iz VII.3. i VII.4.

7. Odredite granice intervala povjerenja pouzdanosti $\gamma = 0{,}95$ za kvantil x_p u neparametarskom modelu, ako je $n = 100$ i:

a) $p = 0{,}05$, b) $p = 0{,}10$, c) $p = 0{,}25$,
d) $p = 0{,}75$, e) $p = 0{,}90$, f) $p = 0{,}95$.

8. Na slučajnom uzorku veličine $n = 10$ izračunajte:

a) $P(Y_3 \leq x_{0,5} \leq Y_8)$,
b) $P(Y_1 \leq x_{0,25} \leq Y_5)$,
c) $P(Y_5 \leq x_{0,75} \leq Y_{10})$,

gdje je $Y_i (i = 1, 3, 5, 8, 10)$ i-ta uređajna statistika, a x_p $(p = 0{,}5\ ;\ 0{,}25\ ;\ 075)$ kvantil p-tog reda.

9. Odredite najmanju moguću veličinu (n) uzorka za koju vrijedi $P(Y_1 \leq M \leq$ $\leq Y_n)$, gdje je Y_1 minimalna, a Y_n maksimalna vrijednost u slučajnom uzorku.

10. Za $n = 20$ odredite kritično područje razine značajnosti $\alpha = 0{,}05$ pri testiranju hipoteze $H_0 : M = 0$, prema alternativnoj hipotezi $H_1 : M \neq 0$, uz primjenu:

a) relacija (31) i (32),
b) relacija (23) i (24),

pri čemu razmotrite pretpostavke da je vrijednost $\widehat{f}_1$ empirijske frekvencije 5, 10 i 15.

11. Nađite kritično područje pri testiranju nul-hipoteze $H_0 : x_{0,25} = x_0$, prema alternativnoj hipotezi $H_1 : x_{0,25} < x_0$, uz $n = 40$, $\alpha = 0{,}10$ i primjenu test-statistike Z_n iz XIV.4.

12. Izvedite formule (33) i (34).

13. Dokažite da Wilcoxonovoj statistici W (formula (38)) pripada ista razdioba vjerojatnosti kao i slučajnoj varijabli V, definiranoj u XIV.5. (formula (39)).

Uputa: Uočite da je svaki član $Z_i R_i$ $(i = 1,\dots,n)$ u (38) jednak jednom i samo jednom od članova $V_1,\dots,V_n$ i pozovite se na činjenice 1–3. u XIV.5.

14. Dopunite tabl. 4. za $n = 5$ i $n = 6$.

15. Izvedite formule (41) i (42).
Uputa: Iskoristite činjenicu da je $\sum_{i=1}^{n} i^2 = \frac{n}{6}(n+1)(2n+1)$.

16. Neka W_1 označuje zbroj rangova negativnih podataka, W_2 zbroj rangova pozitivnih podataka i W Wilcoxonovu statistiku. Dokažite da vrijedi:

a) $W = W_2 - W_1$,

b) $W = \dfrac{n(n+1)}{2} - 2W_1$,

c) $W = 2W_2 - \dfrac{n(n+1)}{2}$.

17. Neka je $\widehat{W}$ test-statistika iz XIV.6. i $m = 2$ i $n = 5$. Odredite skup $A_{\widehat{W}}$ i $\mathrm{P}(\widehat{W} = k)$, $k \in A_{\widehat{W}}$.
Uputa: Uočite da je vjerojatnost događaja $\{\widehat{W} = k\}$ jednaka vjerojatnosti da, nasumce izvlačeći m brojeva iz skupa $\{1, 2, \dots, m+n\}$, dobijete broj k kao vrijednost njihova zbroja.

18. Neka je $\widehat{W}$ test-statistika iz XIV.6. i $m = n = 10$. Primjenom formula (49) i (50) odredite c_1 i c_2 tako da vrijedi $\mathrm{P}(\widehat{W} \le c_1) = \mathrm{P}(\widehat{W} \ge c_2) = 0{,}05$. Usporedite dobiveni rezultat s odgovarajućim vrijednostima u tabl. X. u Dodatku.

19. Dokažite formulu (54).
Uputa: Uočite da $\mathrm{P}(V = v)$ označuje također i vjerojatnost da nasumce izvlačeći k kuglica iz vrećice, koja sadrži m plavih i n bijelih kuglica, dobijemo među njima v plavih kuglica.

20. Izvedite formule za c_1 i c_2 iz (55), uz pretpostavku da $V \sim \mathrm{N}(\mu, \sigma^2)$, gdje su:

a) μ i σ^2 određeni formulama (56),

b) μ i σ^2 određeni formulama (60).

21. Izvedite formule za $\mathrm{E}[V]$ i $\mathrm{D}[V]$ iz (56) za $k = \dfrac{m+n-1}{2}$ i pokažite da se za velike m i n može uzeti da približno vrijedi $\mathrm{E}[V] = \dfrac{m}{2}$, $\mathrm{D}[V] = \dfrac{mn}{4(m+n)}$.

22. Izračunajte $\mathrm{P}(V = 2)$ i $\mathrm{P}(V = 7)$, prema formuli (54), za slučaj $m = 9$ i $n = 11$, te primijenite dobiveni rezultat na problem iz 6. primjera.

23. Izvedite formulu (59).
Uputa: Vidite uputu uz 19. zadatak.

24. Izvedite formulu (61).
Uputa: Primijenite klasičnu definiciju vjerojatnosti u situaciji kada se skup svih mogućih ishoda sastoji od svih mogućih izbora m pozicija za iksove od ukupno raspoloživih $m + n$ pozicija.

DODATAK

TABLICA I.

Vjerojatnosti $\mathrm{P}(X \leq r) = \sum_{i=1}^{r} p_i$ $(r = 0, 1, \ldots, m)$ u binomnoj razdiobi $\mathrm{B}(m, p)$

		p									
m	r	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90
5	0	0,5905	0,3277	0,2373	0,1681	0,0778	0,0312	0,0102	0,0024	0,0003	0,0000
	1	0,9185	0,7373	0,6328	0,5282	0,3370	0,1875	0,0870	0,0308	0,0067	0,0005
	2	0,9914	0,9421	0,8965	0,8369	0,6826	0,5000	0,3174	0,1631	0,0579	0,0086
	3	1,0000	0,9933	0,9844	0,9692	0,9130	0,8125	0,6630	0,4718	0,2627	0,0815
	4	1,0000	0,9997	0,9990	0,9976	0,9898	0,9688	0,9222	0,8319	0,6723	0,4095
	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
10	0	0,3487	0,1074	0,0563	0,0282	0,0060	0,0010	0,0001	0,0000	0,0000	0,0000
	1	0,7361	0,3758	0,2440	0,1493	0,0464	0,0107	0,0017	0,0001	0,0000	0,0000
	2	0,9298	0,6778	0,5256	0,3828	0,1673	0,0547	0,0123	0,0016	0,0001	0,0000
	3	0,9872	0,8791	0,7759	0,6496	0,3823	0,1719	0,0548	0,0106	0,0009	0,0000
	4	0,9984	0,9672	0,9219	0,8497	0,6331	0,3770	0,1662	0,0474	0,0064	0,0002
	5	0,9999	0,9936	0,9803	0,9527	0,8338	0,6230	0,3669	0,1503	0,0328	0,0016
	6	1,0000	0,9999	0,9965	0,9894	0,9452	0,8281	0,6177	0,3504	0,1209	0,0128
	7	1,0000	1,0000	0,9996	0,9984	0,9877	0,9453	0,8327	0,6172	0,3200	0,0702
	8	1,0000	1,0000	1,0000	0,9990	0,9983	0,9893	0,9536	0,8507	0,2642	0,2639
	9	1,0000	1,0000	1,0000	1,0000	0,9999	0,9990	0,9940	0,9718	0,8926	0,6513
	10	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
15	0	0,2059	0,0352	0,0134	0,0047	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,5490	0,1671	0,0802	0,0353	0,0052	0,0005	0,0000	0,0000	0,0000	0,0000
	2	0,8159	0,3980	0,2361	0,1268	0,0271	0,0037	0,0003	0,0000	0,0000	0,0000
	3	0,9444	0,6482	0,4613	0,2969	0,0905	0,0176	0,0019	0,0001	0,0000	0,0000
	4	0,9873	0,8358	0,6865	0,5155	0,2173	0,0592	0,0094	0,0007	0,0000	0,0000
	5	0,9978	0,9389	0,8516	0,7216	0,4032	0,1509	0,0338	0,0037	0,0001	0,0000
	6	0,9997	0,9819	0,9434	0,8689	0,6098	0,3036	0,0951	0,0152	0,0008	0,0000
	7	1,0000	0,9958	0,9827	0,9500	0,7869	0,5000	0,2131	0,0500	0,0042	0,0000
	8	1,0000	0,9999	0,9958	0,9848	0,9050	0,6964	0,3902	0,1311	0,0181	0,0003
	9	1,0000	1,0000	0,9992	0,9963	0,9662	0,8491	0,5968	0,2784	0,0611	0,0023
	10	1,0000	1,0000	0,9999	0,9993	0,9907	0,9408	0,7827	0,4845	0,1642	0,0127
	11	1,0000	1,0000	1,0000	0,9999	0,9981	0,9824	0,9095	0,7031	0,3518	0,0556
	12	1,0000	1,0000	1,0000	1,0000	0,9997	0,9963	0,9729	0,8732	0,6020	0,1841
	13	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9948	0,9647	0,8329	0,4510
	14	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9953	0,9648	0,7941
	15	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	0	0,1216	0,0115	0,0032	0,0008	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,3917	0,0692	0,0243	0,0076	0,0005	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,6769	0,2061	0,0913	0,0355	0,0036	0,0002	0,0000	0,0000	0,0000	0,0000
	3	0,8670	0,4114	0,2252	0,1071	0,0160	0,0013	0,0001	0,0000	0,0000	0,0000
	4	0,9568	0,6296	0,4148	0,2375	0,0510	0,0059	0,0003	0,0000	0,0000	0,0000
	5	0,9887	0,8042	0,6172	0,4164	0,1256	0,0207	0,0016	0,0000	0,0000	0,0000
	6	0,9976	0,9133	0,7858	0,6080	0,2500	0,0577	0,0065	0,0003	0,0000	0,0000
	7	0,9996	0,9679	0,8982	0,7723	0,4159	0,1316	0,0210	0,0013	0,0000	0,0000
	8	0,9999	0,9900	0,9591	0,8867	0,5956	0,2517	0,0565	0,0051	0,0001	0,0000
	9	1,0000	0,9974	0,9861	0,9520	0,7553	0,4119	0,1275	0,0171	0,0006	0,0000
	10	1,0000	0,9994	0,9961	0,9829	0,8725	0,5881	0,2447	0,0480	0,0026	0,0000
	11	1,0000	0,9999	0,9991	0,9949	0,9435	0,7483	0,4044	0,1133	0,0100	0,0001
	12	1,0000	1,0000	0,9998	0,9987	0,9790	0,8684	0,5841	0,2277	0,0321	0,0004
	13	1,0000	1,0000	1,0000	0,9997	0,9935	0,9423	0,7500	0,3920	0,0867	0,0024
	14	1,0000	1,0000	1,0000	1,0000	0,9984	0,9793	0,8744	0,5836	0,1958	0,0113
	15	1,0000	1,0000	1,0000	1,0000	0,9997	0,9941	0,9094	0,7625	0,3704	0,0432
	16	1,0000	1,0000	1,0000	1,0000	1,0000	0,9987	0,9840	0,8929	0,5886	0,1330
	17	1,0000	1,0000	1,0000	1,0000	1,0000	0,9998	0,9964	0,9645	0,7939	0,3231
	18	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9995	0,9924	0,9308	0,6083
	19	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0,9992	0,9885	0,8784
	20	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

TABLICA II.

Vjerojatnosti $P(X \leq r) = \sum_{i=1}^{r} p_i$ $(r = 0, 1, 2, \ldots)$ u Poissonovoj razdiobi $Po(\lambda)$

	λ								
r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0	0,9048	0,8187	0,7408	0,6730	0,6065	0,5488	0,4966	0,4493	0,4066
1	0,9953	0,9825	0,9631	0,9384	0,9098	0,8781	0,8442	0,8088	0,7725
2	0,9998	0,9989	0,9964	0,9921	0,9856	0,9769	0,9659	0,9526	0,9371
3	1,0000	0,9999	0,9997	0,9992	0,9982	0,9966	0,9942	0,9909	0,9865
4		1,0000	1,0000	0,9999	0,9998	0,9996	0,9992	0,9986	0,9977
5				1,0000	1,0000	1,0000	0,9999	0,9998	0,9997
6							1,0000	1,0000	1,0000

	λ								
r	1,0	1,5	2,0	2,5	3,0	3,5	4,0	4,5	5,0
0	0,3679	0,2231	0,1353	0,0821	0,498	0,0302	0,0183	0,0111	0,0067
1	0,7358	0,5578	0,4060	0,2873	0,1991	0,1359	0,0916	0,0611	0,0404
2	0,9197	0,8088	0,6767	0,5438	0,4232	0,3208	0,2381	0,1736	0,1247
3	0,9810	0,9344	0,8571	0,7576	0,6472	0,5366	0,4335	0,3423	0,2650
4	0,9963	0,9814	0,9473	0,8912	0,8153	0,7254	0,6288	0,5321	0,4405
5	0,9994	0,9955	0,9834	0,9580	0,9161	0,8576	0,7851	0,7029	0,6160
6	0,9999	0,9991	0,9955	0,9858	0,9665	0,9347	0,8893	0,8311	0,7622
7	1,0000	0,9998	0,9989	0,9958	0,9881	0,9733	0,9489	0,9134	0,8666
8		1,0000	0,9998	0,9989	0,9962	0,9901	0,9786	0,9597	0,9319
9			1,0000	0,9997	0,9989	0,9967	0,9919	0,9829	0,9682
10				0,9999	0,9997	0,9990	0,9972	0,9933	0,9863
11				1,0000	0,9999	0,9997	0,9991	0,9976	0,9945
12					1,0000	0,9999	0,9997	0,9992	0,9980
13						1,0000	0,9999	0,9997	0,9993
14							1,0000	0,9999	0,9998
15								1,0000	0,9999
16									1,0000

	λ								
r	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0	9,5
0	0,0041	0,0025	0,0015	0,0009	0,0006	0,0003	0,0002	0,0001	0,0001
1	0,0266	0,0174	0,0113	0,0073	0,0047	0,0030	0,0019	0,0012	0,0008
2	0,0884	0,0620	0,0430	0,0296	0,0203	0,0138	0,0093	0,0062	0,0042
3	0,2017	0,1512	0,1118	0,0818	0,0591	0,0424	0,0301	0,0212	0,0149
4	0,3575	0,2851	0,2337	0,1730	0,1321	0,0996	0,0744	0,0550	0,0403
5	0,5289	0,4457	0,3690	0,3007	0,2414	0,1912	0,1496	0,1157	0,0885
6	0,6860	0,6063	0,5265	0,4497	0,3782	0,3134	0,2562	0,2068	0,1649
7	0,8095	0,7440	0,6728	0,5987	0,5246	0,4530	0,3856	0,3239	0,2687
8	0,8944	0,8472	0,7916	0,7291	0,6620	0,5925	0,5231	0,4557	0,3918
9	0,9462	0,9161	0,8774	0,8305	0,7764	0,7166	0,6530	0,5874	0,5218
10	0,9747	0,9574	0,9332	0,9015	0,8622	0,8159	0,7634	0,7060	0,6453
11	0,9890	0,9799	0,9661	0,9466	0,9208	0,8881	0,8487	0,8030	0,7520
12	0,9955	0,9912	0,9840	0,9730	0,9573	0,9362	0,9091	0,8758	0,8364
13	0,9983	0,9964	0,9929	0,9872	0,9784	0,9658	0,9486	0,9261	0,8981
14	0,9994	0,9986	0,9970	0,9943	0,9897	0,9827	0,9726	0,9585	0,9400
15	0,9998	0,9995	0,9988	0,9976	0,9954	0,9918	0,9862	0,9780	0,9665
16	1,0000	0,9996	0,9996	0,9990	0,9980	0,9963	0,9934	0,9889	0,9823
17	1,0000	0,9999	0,9998	0,9996	0,9992	0,9984	0,9970	0,9947	0,9911
18		1,0000	0,9999	0,9999	0,9997	0,9994	0,9987	0,9976	0,9957
19			1,0000	1,0000	0,9999	0,9997	0,9995	0,9989	0,9980
20					1,0000	0,9999	0,9998	0,9996	0,9991
21						1,0000	0,9999	0,9998	0,9996
22							1,0000	0,9999	0,9999
23								1,0000	0,9999
24									1,0000

TABLICA II. (nastavak)

	λ								
r	10,0	11,0	12,0	13,0	14,0	15,0	16,0	17,0	18,0
0	0,0000	0,0000	0,0000						
1	0,0005	0,0002	0,0001	0,0000					
2	0,0028	0,0012	0,0005	0,0002	0,0000				
3	0,0103	0,0049	0,0023	0,0010	0,0001	0,0000	0,0000		
4	0,0293	0,0151	0,0076	0,0037	0,0005	0,0002	0,0001	0,0000	0,0000
5	0,0671	0,0375	0,0203	0,0107	0,0018	0,0009	0,0004	0,0002	0,0001
6	0,1301	0,0786	0,0458	0,0259	0,0055	0,0028	0,0014	0,0007	0,0003
7	0,2202	0,1432	0,0895	0,0540	0,0142	0,0076	0,0040	0,0021	0,0010
8	0,3328	0,2320	0,1550	0,0998	0,0316	0,0180	0,0100	0,0054	0,0029
9	0,4579	0,3405	0,2424	0,1658	0,0621	0,0374	0,0220	0,0126	0,0071
10	0,5830	0,4599	0,3472	0,2517	0,1094	0,0699	0,0433	0,0261	0,0154
11	0,6968	0,5793	0,4616	0,3532	0,1757	0,1185	0,0774	0,0491	0,0304
12	0,7916	0,6887	0,5760	0,4631	0,2600	0,1848	0,1270	0,0847	0,0549
13	0,8645	0,7813	0,6815	0,5730	0,3585	0,2676	0,1931	0,1350	0,0917
14	0,9165	0,8540	0,7720	0,6751	0,4664	0,3632	0,2745	0,2009	0,1426
15	0,9513	0,9074	0,8444	0,7636	0,5704	0,4657	0,3675	0,2808	0,2081
16	0,9730	0,9441	0,8987	0,8355	0,6694	0,5681	0,4667	0,3715	0,2867
17	0,9857	0,9678	0,9370	0,8905	0,7559	0,6641	0,5660	0,4677	0,3750
18	0,9928	0,9823	0,9626	0,9302	0,8272	0,7489	0,6593	0,5640	0,4686
19	0,9965	0,9907	0,9787	0,9573	0,8826	0,8195	0,7423	0,6550	0,5622
20	0,9984	0,9953	0,9884	0,9750	0,9235	0,8752	0,8122	0,7363	0,6509
21	0,9993	0,9977	0,9939	0,9859	0,9521	0,9170	0,8682	0,8055	0,7307
22	0,9997	0,9990	0,9970	0,9924	0,9712	0,9469	0,9108	0,8615	0,7991
23	0,9999	0,9995	0,9985	0,9960	0,9833	0,9673	0,9418	0,9047	0,8581
24	1,0000	0,9998	0,9993	0,9980	0,9907	0,9805	0,9633	0,9367	0,8989
25		0,9999	0,9997	0,9990	0,9950	0,9888	0,9777	0,9594	0,9317
26		1,0000	0,9999	0,9995	0,9974	0,9938	0,9869	0,9748	0,9554
27			0,9999	0,9998	0,9987	0,9967	0,9925	0,9848	0,9718
28			1,0000	0,9999	0,9994	0,9983	0,9959	0,9912	0,9827
29				1,0000	0,9997	0,9991	0,9978	0,9950	0,9897
30					0,9999	0,9996	0,9989	0,9973	0,9941
31					0,9999	0,9998	0,9994	0,9986	0,9967
32					1,0000	0,9999	0,9997	0,9993	0,9982
33						1,0000	0,9999	0,9996	0,9990
34							0,9999	0,9998	0,9995
35							1,0000	0,9999	0,9998
36								1,0000	0,9999
37									0,9999
									1,0000

TABLICA III.

Vjerojatnosti $P(Z \leq z) = \Phi(z)$ u standardnoj normalnoj razdiobi $N(0,1)$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0352	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0722	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

TABLICA III. (nastavak)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9278	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

TABLICA IV.

Vrijednosti gama–funkcije $\Gamma(\lambda)$, $\lambda > 0$

λ	$\Gamma(\lambda)$	λ	$\Gamma(\lambda)$	λ	$\Gamma(\lambda)$	λ	$\Gamma(\lambda)$
1,01	0,9943	1,26	0,9044	1,51	0,8866	1,76	0,9214
1,02	0,9888	1,27	0,9025	1,52	0,8870	1,77	0,9238
1,03	0,9835	1,28	0,9007	1,53	0,8876	1,78	0,9262
1,04	0,9784	1,29	0,8990	1,54	0,8882	1,79	0,9288
1,05	0,9735	1,30	0,8975	1,55	0,8889	1,80	0,9314
1,06	0,9687	1,31	0,8960	1,56	0,8896	1,81	0,9341
1,07	0,9642	1,32	0,8946	1,57	0,8905	1,82	0,9368
1,08	0,9597	1,33	0,8934	1,58	0,8914	1,83	0,9397
1,09	0,9555	1,34	0,8922	1,59	0,8924	1,84	0,9426
1,10	0,9514	1,35	0,8912	1,60	0,8935	1,85	0,9456
1,11	0,9474	1,36	0,8902	1,61	0,8947	1,86	0,9487
1,12	0,9436	1,37	0,8893	1,62	0,8959	1,87	0,9518
1,13	0,9399	1,38	0,8885	1,63	0,8972	1,88	0,9551
1,14	0,9364	1,39	0,8879	1,64	0,8986	1,89	0,9584
1,15	0,9330	1,40	0,8873	1,65	0,9001	1,90	0,9618
1,16	0,9298	1,41	0,8868	1,66	0,9017	1,91	0,9652
1,17	0,9267	1,42	0,8864	1,67	0,9033	1,92	0,9688
1,18	0,9237	1,43	0,8860	1,68	0,9050	1,93	0,9724
1,19	0,9209	1,44	0,8858	1,69	0,9068	1,94	0,9761
1,20	0,9182	1,45	0,8857	1,70	0,9086	1,95	0,9799
1,21	0,9156	1,46	0,8856	1,71	0,9106	1,96	0,9837
1,22	0,9131	1,47	0,8856	1,72	0,9126	1,97	0,9877
1,23	0,9108	1,48	0,8857	1,73	0,9147	1,98	0,9917
1,24	0,9085	1,49	0,8859	1,74	0,9168	1,99	0,9958
1,25	0,9064	1,50	0,8862	1,75	0,9191	2,00	1,0000

Za izračunavanje vrijednosti izvan intervala $\langle 1, 2]$ primijenite formulu $\Gamma(\lambda) = (\lambda - 1)\Gamma(\lambda - 1)$.

TABLICA V.

Vrijednosti $G_n^{-1}(p)$. (G_n – f.r.v. za Studentovu razdiobu sa n stupnjeva slobode.)

n	p				
	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,669	2,045	2,462	2,756
∞	1,282	1,645	1,960	2,326	2,576

TABLICA VI.

Vrijednosti $H_n^{-1}(p)$. (H_n – f.r.v. za hikvadrat–razdiobu sa n stupnjeva slobode.)

n	p = 0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
1	0,0000	0,0001	0,0010	0,004	3,841	5,024	6,635	7,879
2	0,0100	0,0201	0,0506	0,103	5,991	7,378	9,210	10,597
3	0,0717	0,115	0,216	0,352	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	10,070	12,832	15,086	16,750
6	0,676	0,872	1,237	1,635	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	24,996	27,488	30,578	32,997
16	5,142	5,812	6,908	7,962	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	36,415	39,364	42,980	45,558
25	10,520	11,524	13,120	14,611	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	40,113	43,194	46,963	49,645
28	12,461	13,565	15,308	16,928	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	43,773	46,979	50,892	53,672

TABLICA VII.

Vrijednosti $F_{rs}^{-1}(p)$. (F_{rs} – f.r.v. za F–razdiobu sa (r, s) stupnjeva slobode.)

$p = 0,95$	r																	
s	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18.51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	7,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3.18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

TABLICA VII. (nastavak)

$p = 0{,}99$	r																		
s	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999,5	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,46	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

TABLICA VIII.

Vrijednosti veličine c_0 u KS-test. (Formula (10) u X. 2.)

n	$\alpha = 0{,}20$	$\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}02$	$\alpha = 0{,}01$
1	0,900	0,950	0,975	0,990	0,995
2	0,684	0,776	0,842	0,900	0,929
3	0,565	0,636	0,708	0,785	0,829
4	0,493	0,565	0,624	0,689	0,734
5	0,447	0,509	0,563	0,627	0,669
6	0,410	0,468	0,519	0,577	0,617
7	0,381	0,436	0,483	0,538	0,576
8	0,359	0,410	0,454	0,507	0,542
9	0,339	0,387	0,430	0,480	0,513
10	0,323	0,369	0,409	0,457	0,486
11	0,308	0,352	0,391	0,437	0,468
12	0,296	0,338	0,375	0,419	0,449
13	0,285	0,325	0,361	0,404	0,432
14	0,275	0,314	0,349	0,390	0,418
15	0,266	0,304	0,338	0,377	0,404
16	0,258	0,295	0,327	0,366	0,392
17	0,250	0,286	0,318	0,355	0,381
18	0,244	0,279	0,309	0,346	0,371
19	0,237	0,271	0,301	0,337	0,361
20	0,232	0,265	0,294	0,329	0,352
21	0,226	0,259	0,287	0,321	0,344
22	0,221	0,253	0,281	0,314	0,337
23	0,216	0,247	0,275	0,307	0,330
24	0,212	0,242	0,269	0,301	0,323
25	0,208	0,238	0,264	0,295	0,317
26	0,204	0,233	0,259	0,290	0,311
27	0,200	0,229	0,254	0,284	0,305
28	0,197	0,225	0,250	0,279	0,300
29	0,193	0,221	0,246	0,275	0,295
30	0,190	0,218	0,242	0,270	0,290
35	0,177	0,202	0,224	0,251	0,269
40	0,165	0,189	0,210	0,235	0,252
45	0,156	0,179	0,198	0,222	0,238
50	0,148	0,170	0,188	0,211	0,226
55	0,142	0,162	0,180	0,201	0,216
60	0,136	0,155	0,172	0,193	0,207
65	0,131	0,149	0,166	0,185	0,199
70	0,126	0,144	0,160	0,179	0,192
75	0,122	0,139	0,154	0,173	0,185
80	0,118	0,135	0,150	0,167	0,179
85	0,114	0,131	0,145	0,162	0,174
90	0,111	0,127	0,141	0,158	0,169
95	0,108	0,124	0,137	0,154	0,165
100	0,106	0,121	0,134	0,150	0,161

TABLICA IX.

Vrijednosti veličine c za koju vrijedi $P(|W| \geq c) = \alpha$. (W - Wilcoxonova statistika XIV.5.)

n	$\alpha = 0{,}05$	$\alpha = 0{,}10$
	c	c
6	21	17
7	24	20
8	28	24
9	33	29
10	39	33
11	44	38
12	50	44
13	57	49
14	63	53
15	70	60
16	76	64
17	83	71
18	91	77
19	98	82
20	106	90

TABLICA X.

Najveći cijeli broj c_1 i najmanji cijeli broj c_2 za koje vrijedi $P(\widehat{W} \leq c_1) = P(\widehat{W} \geq c_2) = \frac{\alpha}{2}$. ($\widehat{W}$ - test statistika MWW-testa, XIV.6.)

m	n	$\alpha = 0,10$		$\alpha = 0,05$		$\alpha = 0,02$		$\alpha = 0,01$	
		c_1	c_2	c_1	c_2	c_1	c_2	c_1	c_2
2	5	3	13						
	6	3	15						
	7	3	17						
	8	4	18	3	19				
	9	4	20	3	21				
	10	4	22	3	23				
3	3	6	15						
	4	6	18						
	5	7	20	6	21				
	6	8	22	7	23				
	7	8	25	7	26	6	27		
	8	9	27	8	28	6	30		
	9	10	29	8	31	7	32	6	33
	10	10	32	9	33	7	35	6	36
4	4	11	25	10	26				
	5	12	28	11	29	10	30		
	6	13	31	12	32	11	33	10	34
	7	14	34	13	35	11	37	10	38
	8	15	37	14	38	12	40	11	41
	9	16	40	15	41	13	43	11	45
	10	17	43	15	45	13	47	12	48
5	5	19	36	17	38	16	39	15	40
	6	20	40	18	42	17	43	16	44
	7	21	44	20	45	18	47	17	48
	8	23	47	21	49	19	51	17	53
	9	24	51	22	53	20	55	18	57
	10	26	54	23	57	21	59	19	61
6	6	28	50	26	52	24	54	23	55
	7	29	55	27	57	25	59	24	60
	8	31	59	29	61	27	63	25	65
	9	33	63	31	65	28	68	26	70
	10	35	67	32	70	29	73	27	75
7	7	39	66	36	69	34	71	32	73
	8	41	71	38	74	36	76	34	78
	9	43	76	40	79	37	82	35	84
	10	45	81	42	84	39	87	37	89
8	8	51	85	49	87	46	90	43	93
	9	54	90	51	93	48	96	45	99
	10	56	96	53	99	50	102	47	105
9	9	66	105	63	108	59	112	56	115
	10	69	111	65	115	61	119	58	122
10	10	82	128	78	132	74	136	71	139

TABLICA XI.

Vjerojatnosti $P(R \leq r)$ za test-statistiku R. (Test serija, XIV. 8.)

	r																		
(m,n)	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(2,3)	0,200	0,500	0,900	1,000															
(2,4)	0,133	0,400	0,800	1,000															
(2,5)	0,095	0,333	0,714	1,000															
(2,6)	0,071	0,286	0,643	1,000															
(2,7)	0,056	0,250	0,583	1,000															
(2,8)	0,044	0,222	0,533	1,000															
(2,9)	0,036	0,200	0,491	1,000															
(2,10)	0,030	0,182	0,455	1,000															
(3,3)	0,100	0,300	0,700	0,900	1,000														
(3,4)	0,057	0,200	0,543	0,800	0,971	1,000													
(3,5)	0,036	0,143	0,429	0,714	0,929	1,000													
(3,6)	0,024	0,107	0,345	0,643	0,881	1,000													
(3,7)	0,017	0,083	0,283	0,583	0,833	1,000													
(3,8)	0,012	0,067	0,236	0,533	0,788	1,000													
(3,9)	0,009	0,055	0,200	0,491	0,745	1,000													
(3,10)	0,007	0,045	0,171	0,455	0,706	1,000													
(4,4)	0,029	0,114	0,371	0,629	0,886	0,971	1,000												
(4,5)	0,016	0,071	0,262	0,500	0,786	0,929	0,992	1,000											
(4,6)	0,010	0,048	0,190	0,405	0,690	0,881	0,976	1,000											
(4,7)	0,006	0,033	0,142	0,333	0,606	0,833	0,954	1,000											
(4,8)	0,004	0,024	0,109	0,279	0,533	0,788	0,929	1,000											
(4,9)	0,003	0,018	0,085	0,236	0,471	0,745	0,902	1,000											
(4,10)	0,002	0,014	0,068	0,203	0,419	0,706	0,874	1,000											

TABLICA XI. (nastavak)

(m, n)	r = 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(5,5)	0,008	0,040	0,167	0,357	0,643	0,833	0,960	0,992	1,000										
(5,6)	0,004	0,024	0,110	0,262	0,522	0,738	0,911	0,976	0,998	1,000									
(5,7)	0,003	0,015	0,076	0,197	0,424	0,652	0,854	0,955	0,992	1,000									
(5,8)	0,002	0,010	0,054	0,152	0,347	0,576	0,793	0,929	0,984	1,000									
(5,9)	0,001	0,007	0,039	0,119	0,287	0,510	0,734	0,902	0,972	1,000									
(5,10)	0,001	0,005	0,029	0,095	0,239	0,455	0,678	0,874	0,958	1,000									
(6,6)	0,002	0,013	0,067	0,175	0,392	0,608	0,825	0,933	0,987	0,998	1,000								
(6,7)	0,001	0,008	0,043	0,121	0,296	0,500	0,733	0,879	0,966	0,992	0,999	1,000							
(6,8)	0,001	0,005	0,028	0,086	0,226	0,413	0,646	0,821	0,937	0,984	0,998	1,000							
(6,9)	0,000	0,003	0,019	0,063	0,175	0,343	0,566	0,762	0,902	0,972	0,994	1,000							
(6,10)	0,000	0,002	0,013	0,047	0,137	0,288	0,497	0,706	0,864	0,958	0,990	1,000							
(7,7)	0,001	0,004	0,025	0,078	0,209	0,383	0,617	0,791	0,922	0,975	0,996	0,999	1,000						
(7,8)	0,000	0,002	0,015	0,051	0,149	0,296	0,514	0,704	0,867	0,949	0,988	0,998	1,000	1,000					
(7,9)	0,000	0,001	0,010	0,035	0,108	0,231	0,427	0,622	0,806	0,916	0,975	0,994	0,999	1,000					
(7,10)	0,000	0,001	0,006	0,024	0,080	0,182	0,355	0,549	0,743	0,879	0,957	0,990	0,998	1,000					
(8,8)	0,000	0,001	0,009	0,032	0,100	0,214	0,405	0,595	0,786	0,900	0,968	0,991	0,999	1,000	1,000				
(8,9)	0,000	0,001	0,005	0,020	0,069	0,157	0,319	0,500	0,702	0,843	0,939	0,980	0,996	0,999	1,000	1,000			
(8,10)	0,000	0,001	0,003	0,013	0,048	0,117	0,251	0,419	0,621	0,782	0,903	0,964	0,990	0,998	1,000	1,000			
(9,9)	0,000	0,000	0,003	0,012	0,044	0,109	0,238	0,399	0,601	0,762	0,891	0,956	0,988	0,997	1,000	1,000	1,000		
(9,10)	0,000	0,000	0,002	0,008	0,029	0,077	0,179	0,319	0,510	0,681	0,834	0,923	0,974	0,992	0,999	1,000	1,000	1,000	
(10,10)	0,000	0,000	0,001	0,004	0,019	0,051	0,128	0,242	0,414	0,586	0,758	0,872	0,949	0,981	0,996	0,999	1,000	1,000	1,000

TABLICA XII.

Vrijednosti veličina c_1 i d_1 u DW–testu

α	n	$r = 1$		$r = 2$		$r = 3$		$r = 4$		$r = 5$	
		c_1	d_1	c_1	d_1	c_1	d_1	c_1	d_1	c_1	d_1
	15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
	16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
	17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
	18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
	19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
	20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
	21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
	22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
	23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
	24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
	25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
	26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
	27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
	28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
	29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
	30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
	31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
	32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
	33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
0,05	34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
	35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
	36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
	37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
	38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
	39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
	40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
	45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
	50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
	55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
	60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
	65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
	70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
	75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
	80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
	85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
	90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
	95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
	100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

TABLICA XII. (nastavak)

α	n	$r = 1$		$r = 2$		$r = 3$		$r = 4$		$r = 5$	
		c_1	d_1	c_1	d_1	c_1	d_1	c_1	d_1	c_1	d_1
	15	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
	16	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
	17	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
	18	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
	19	0,93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
	20	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
	21	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
	22	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
	23	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
	24	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
	25	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
	26	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
	27	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
	28	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
	29	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
	30	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
	31	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
	32	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
	33	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
0,01	34	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
	35	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
	36	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
	37	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
	38	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
	39	1,24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
	40	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
	45	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
	50	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
	55	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
	60	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
	65	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
	70	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
	75	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
	80	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
	85	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
	90	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
	95	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
	100	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65

Popis literature

[1] Anderson, R.L., Bancroft, T.A., *Statistical Theory in Research*, McGraw-Hill, 1952.

[2] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*, J. Wiley, 1958.

[3] Benjamin, J.R., Cornell, C.A., *Probability, Statistics and Decision for Civil Engineers*, McGraw-Hill, 1970.

[4] Breiman, L., *Statistics With a View Toward Applications*, Houghton Mifflin Company, Boston, 1973.

[5] Cody, R.P., Smith, J.K., *Applied Statistics and the SAS Programming Language*, North-Holland, Amsterdam, 1987.

[6] Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, 1961.

[7] Dixon, W.J., Massey, F.J., *Introduction to Statistical Analysis*, McGraw-Hill, 1957.

[8] Draper, N., Smith, H., *Applied Regression Analysis*, J. Wiley, 1966.

[9] Dunin-Barkovskij, I.V., Smirnov, N.V., *Teorija vjerojatnostej i matematičeskaja statistika v tehnike*, Gostehizdat, 1955.

[10] Durbin, J., Watson, G.S.,*"Testing for Serial Correlation in Last Squares Regression"*, Biometrika, 37, 409–428, 1950, 38, 159–178, 1951.

[11] Feller, W., *An Introduction to Probability Theory and Its Applications*, J. Wiley, Vol.I. 1961, Vol.II. 1966.

[12] Fisher, R.A., *"On the mathematical foundations of theoretical statistics"*, Phil. Trans. Roy. Soc. London (A), 222, 309–368, 1922.

[13] Fisher, R.A., *Contributions to Mathematical Statistics*, Editor W.A. Shewhart, J. Wiley, 1950.

[14] Fisher, R.A., *Statistical Methods for Research Workers*, Oliver and Boyd, Edinbourgh, 1958.

[15] Galambos, J., *The Asymptotic Theory of Extreme Order Statistics*, J. Wiley, 1978.

[16] Gumbel, E.J., *Statistics of Extremes*, Columbia University Press, 1958.

[17] Hald, A., *Statistical Tables and Formulas*, J. Wiley, 1962.

[18] Hogg, R.V., *"Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory"*, J. Amer. Stat. Assoc. 69, 909, 1974.

[19] Hogg, R.V., Craig, A.T., *Introduction to Mathematical Statistics*, Macmillan Publishing Co. Inc., 1978.

[20] Huber, P., *"Robust Statistics: A Review"*, Annals Math. Stat., 43, 1041, 1972.

[21] Ivković, Z.A., *Matematička statistika*, Naučna knjiga, Beograd, 1976.

[22] Jamnik, R., *Matematična statistika*, Državna založba Slovenije, Ljubljana, 1980.

[23] Kreyszig, E., *Introductory Mathematical Statistics*, J. Wiley, 1970.

[24] Kullback, S., *Information Theory and Statistics*, J. Wiley, 1959.

[25] Lehmann, E.L., *Testing Statistical Hypotheses*, J. Wiley, 1959.

[26] Loève, M., *Probability Theory*, D. Van Nostrand Company Inc., 1960.

[27] Mandel, J., *The Statistical Analysis of Experimental Data*, Interscience Publishers, 1964.

[28] Mann, H.B., Whitney, D.R., *"On a test of whether one or two random variables is stochastically larger than the other"*, Annals Math. Stat. 18, 50–60, 1947.

[29] Mathisen, H.C., *"A method of testing the hypothesis that two samples are from the same population"*, Annals Math. Stat. 14, 188–194, 1943.

[30] Neyman, J., *"On the problem of confidence intervals"*, Annals Math. Stat. 6, 111–116, 1935.

[31] Neyman, J., Pearson, E.S., *"On the use and interpretation of certain test criteria for purposes of statistical inference"*, Biometrika, 20A, 175–240 and 263–249, 1928.

[32] Neyman, J., Pearson, E.S., *"On the problem of the most efficient tests of statistical hypotheses"*, Phil. Trans. Roy. Soc. London (A) 231, 289–337, 1933.

[33] Newbold, P., *Statatistics for Business and Economics*, Prentice-Hall International Inc., 1988.

[34] Pauše, Ž., *Vjerojatnost, informacija, stohastički procesi*, Školska knjiga, Zagreb, 1988.

[35] Pavlić, I., *Statistička teorija i primjena*, Tehnička knjiga, Zagreb, 1985.

[36] Pearson, K., *"On the criterion that a system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling"*, Phil. Mag. V, 50, 157–175, 1900.

[37] Rènyi, A., *Probability Theory*, North Holland Publishing Co., 1970.

[38] Sarapa, N., *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 1988.

[39] Savage, L.J., *The Foundations of Statistics*, J. Wiley, 1954.

[40] Scheffè, H., *The Analysis of Variance*, J. Wiley, 1963.

[41] Schmetterer, L., *Einführung in die mathematische Statistik*, Springer-Verlag, 1966.

[42] Serdar, V., Šošić, I., *Uvod u statistiku*, Školska knjiga, Zagreb, 1986.

[43] Smirnov, N., *"On the estimation of the discrepancy between empirical curves of distribution for two independent samples"*, Bull. Math. Univ. Moscow, 2(2), 3–14, 1939.

[44] Smith, G.N., *Probability and Statistics in Civil Engineering*, Collins, London, 1986.

[45] Srivastava, M.S., Carter, E.M., *An Introduction to Applied Multivariate Statistics*, North-Holland, Amsterdam, 1983.

[46] "Student", *"The probable error of mean"*, Biometrika, 6, 1–25, 1908.

[47] Ugrin-Šparac, D., *Primijenjena teorija vjerojatnosti*, Sveučilišna naklada Liber, Zagreb, 1975.

[48] Ugrin-Šparac, G., *Neke metode istraživanja generatora pseudoslučajnih brojeva*, Doktorska disertacija, Sveučilište u Zagrebu, 1986.

[49] Vranić, V., *Vjerojatnost i statistika*, Tehnička knjiga, Zagreb, 1971.

[50] Wald, A., *Asymptotically shortest confidence intervals*, Annals Math. Stat., 13, 127–137, 1942.

[51] Wald, A., *Statistical Decision Functions*, J. Wiley, 1950.

[52] Wald, A., *Sequential Analysis*, J. Wiley, 1974.

[53] Walpole, R.E., Myers, R.H., *Probability and Statistics for Engineers and Scientists*, Macmillan, 1978.

[54] Wilks, S.S., *Mathematical Statistics*, J. Wiley, 1962.

[55] Williams, E.J., *Regression Analysis*, J. Wiley, 1959.

[56] Wonnacott, T.H., Wonnacott, R.J., *Introductory Statistics*, J. Wiley, 1969.

[57] Zacks, S., *The Theory of Statistical Inference*, J. Wiley, 1971.

Kazalo

N

O

P

R

S

T

U

V

W